National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

THE LOGIC OF SELF-EFFACEMENT

by

Paul Viminitz ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Philosophy

Edmonton, Alberta
Fall, 1995

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

THE AUTHOR HAS GRANTED AN IRREVOCABLE NON-EXCLUSIVE LICENCE ALLOWING THE NATIONAL LIBRARY OF CANADA TO REPRODUCE, LOAN, DISTRIBUTE OR SELL COPIES OF HIS/HER THESIS BY ANY MEANS AND IN ANY FORM OR FORMAT, MAKING THIS THESIS AVAILABLE TO INTERESTED PERSONS.

L'AUTEUR A ACCORDE UNE LICENCE IRREVOCABLE ET NON EXCLUSIVE PERMETTANT A LA BIBLIOTHEQUE NATIONALE DU CANADA DE REPRODUIRE, PRETER, DISTRIBUER OU VENDRE DES COPIES DE SA THESE DE QUELQUE MANIERE ET SOUS QUELQUE FORME QUE CE SOIT POUR METTRE DES EXEMPLAIRES DE CETTE THESE A LA DISPOSITION DES PERSONNE INTERESSEES.

THE AUTHOR RETAINS OWNERSHIP OF THE COPYRIGHT IN HIS/HER THESIS. NEITHER THE THESIS NOR SUBSTANTIAL EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT HIS/HER PERMISSION.

L'AUTEUR CONSERVE LA PROPRIETE DU DROIT D'AUTEUR QUI PROTEGE SA THESE. NI LA THESE NI DES EXTRAITS SUBSTANTIELS DE CELLE-CI NE DOIVENT ETRE IMPRIMES OU AUTREMENT REPRODUITS SANS SON AUTORISATION.

ISBN 0-612-06305-4

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Paul Kenneth Viminitz

TITLE OF THESIS: The Logic of Self-Effacement

DEGREE: Doctor of Philosophy

YEAR THIS DEGREE GRANTED: 1995

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Box 55, Rockton, Ontario, L0R 1X0

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled THE LOGIC OF SELF-EFFACEMENT here submitted by PAUL KENNETH VIMINITZ in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Wes Cooper

Dr. Wesley E. Cooper, Supervisor

Allen Carlson

Dr. Allen Carlson

Bruce Hunter

Dr. Bruce Hunter

John A Baker

Dr. John Baker

Dr. Don Carmichael

Peter Danielson

Dr. Peter Danielson

November 18, 1994

This work is dedicated to the memory of
those with whom God broke covenant ...
and who coped as best they could.

Abstract

In Morals by Agreement, David Gauthier tells what I take to be the most convincing story yet about what it is we might be doing when we are thinking, talking, and acting morally, a story that can be extended, I submit, to the more general case of acting defensibly. I begin this enquiry, then, with a largely supportive rehearsal of Gauthier's story, including its domain of discourse, its motivation, its methodology, its conception of rationality and, finally, its pivotal move.

What shall be of especial interest, however, is less the substantive details of Gauthier's account than its explanatory schema, the central feature of which being that there are certain 'games' - games paradigmatic of the human interpersonal condition - the winning strategy for which in-volves the performance of an operation on some feature of one's own psychology. Of these the most interesting will prove to be what might be called 'second-order' self-modifications, that is, operations performed on one's own psychology that are designed to ensure the development of - and/or once developed, the protection of - some of these aforementioned first-order modifications.

But of these, in turn, I will be focussing on second-order self-modifications designed solely to ensure the unreadoptability of some feature of the original psychology. I call these features - that is, those it behooves us to render unrecoverable - 'self-effacing'. And my claim is that a comprehensive analysis of the rationality of self-effacement can make good on a promise of incalculable explanatory power.

In the course of this analysis, however, it will emerge that there is a paradox within the rationality of self-effacement, the resolution of which may require a modest revision - but a revision nonetheless - to the conception of rationality with which we started out. This will thrust us squarely into the longstanding debate - only now armed with an agenda of our own - over whether prudence is to be characterized strictly in terms of current preferences which are simply higher-ordered, or rather in terms of first-order preferences which can also be anticipated and/or remembered.

I conclude (and also intersperse) with a series of comments of what ramifications, if any, our findings might have: for the completion of the meta-ethical program with which we (and Gauthier) started out, for the

debate over the conditions on moral considerability, for naturalized epistemology - and, therein, for the residual problem of skepticism - and, finally, for the possibility of yet another convincing story about what it is we might be doing when we are thinking, talking and doing philosophy itself.

Acknowledgements

I would like to thank the following people who, wittingly or un-, contributed immeasureably to my thinking about this project, and my finally completing it: Drs. Richmond Campbell, Robert Martin and Duncan MacIntosh, all at Dalhousie University, Drs. Wes Cooper, Bruce Hunter, Allen Carlson and Martin Tweedale, at the University of Alberta, Dr. John Baker at the University of Calgary, Dr. Peter Danielson at the University of British Columbia, and Dr. Elisabeth Boetzkes at McMaster University. Martin had only to prod; Wes only to prod and supervise; Bruce, Allen, John and Peter only to read. And, probably, Rich, Bob and Duncan knew not even what they had wrought. But Elisabeth, being not only my intellectual confidante and provocateur but, more onerous for her yet, my wife, bore the lioness' share of the burden of my burdens - and theirs. To her especially, ta!

Table of Contents

Introduction and Roadmap 1

1. The Man in the Glass Booth

The Man in the Glass Booth 4
1. Questions 6
2. Reactions 8
3. Personal Identity 11
4. Crimes Against the Self 26
5. Defensible Self-Projection 30
6. Gauthier 35

2. The Contractarian Schema

1 The Schema 43
2. Inter-Level Reduction -
Objections and Replies 44
3. The Prisoners' Dilemma 50
4. The Logic of Leviathan 53
5. The Role of the Sovereign 54
6. The Corruption Problem 56
7. The Expense Problem 59
8. The Compliance Problem 61
9. The Empirical Inadequacy Problem 69

3. The Logic of Morals by Agreement

1. Constrained Maximization 72
2. The Incoherence and Undermotivation Problems .. 73
3. Danielson's *Artificial Morality* 74

4. Campbell's and Danielson's Solutions to
the Incoherence Problem 78
5. Danielson's and Campbell's Solutions to
the Incoherence Problem Revisited 81
6. Gauthier's Solution to
the Compliance Problem 84
7. The Rationale Behind
the Re-Definition of Superiority 87
8. Danielson's Solution to
the Undermotivation Problem 91
9. The Implications of Prudential Superiority 93
10. The Definition of Superiority Revisited and,
Thereby, a Revisit Too to Danielson's
Solution to the Undermotivation Problem .. 96
11. Constrained Maximation Repaired 101
12. Taking the 'Pill' 102
13. Translucency 104
14. The Domain of Discourse 106
15. The Logic of Precommitment in General 111
16. Precommitment and Decidability 123

4. Prisoners' Dilemma is Not a Newcomb's Problem

1. Background 129
2. Equivalence 131
3. Equivalence and Deconstruction 133
4. Deconstructionism,
a.k.a. the Symmetry Argument 134

5. Why the NP *Is* Deconstructable 137

6. Equivalence and Independence 139

7. Some Side-Issues 142

8. Can the PD be Deconstructed
in Its Own Right? 143

9. How to Play Chicken 145

10. How to Win at Chicken 151

11. Prisoners' Dilemma is a Game of Chicken 154

12. Asymmetry and Mixed Strategies 157

13. Independence Revisited 161

5. Rationality

1. Why Need Our Schema Be So Broad? 164

2. Mentality and Preferences 166

3. Freedom 173

4. Information 176

5. Interactivity 180

6. The Marxist/Feminist Objection
- The Underdetermination Problem 181

7. The Existentialist/Postmodernist Objection
- The Privileged Perspective Problem 189

8. The Postmodernist/Deconstructionist Objection
- The Prescriptivity Problem 199

6. The Moral Dialectic

1. Dispositional Pluralism 222

2. The Transparency/Translucency Problem 227

3. Combatting Pluralism 231

4. Combatting Recidivism 237
5. The Empirical Inadequacy of First-Order Moves Alone 241
6. (N>1)-Order Moves: Their Depth 244
7. Their Range 248
8. And Their Flexibility 251

7. The Logic of Self-Effacement

1. Changing One's Mind (The Story So Far) 252
2. The Effacement/Self-Effacement Distinction 258
3. Self-Effacement/Replacement Defined 261
4. Distinctions Within Self-Effacement - Strong/Weak 264
5. Symmetrical/Asymmetrical 273
6. But First an Aside About Stability 274
7. Origins 277
8. The Force of the 'May' 281
9. Residual Problems 285
10. Defaults 290
11. Contextual/Categorical 293
12. Mediated/Direct 295
13. Self-Effacement and Its Cognates - Self-Defeat 297
14. Self-Deception 300

8. Resolving the Paradoxes of Self-Effacement

1. Choices - Hers, His, and Mine 307
2. An Interim Summary 314

3. Our Options 315
4. The MacIntosh Gloss on Prudence Theories 319
5. Reservations 331
6. The 'Meta-' Move 334
7. What Decides Between the 'Correct' Theory and the Meta-Theory? 339
8. How to Be - and Suffer Being - One's Own Other 343
9. The Prescriptivity Problem Revisited 350
10. Personal Identity Revisited 354
11. Epistemology on a Need-to-Know-Only Basis 356
12. Theodicy 359
13. Might the Logic of Self-Effacement be Self-Effacing? 360
Notes to "Introduction and Roadmap" and to Chapter 1 364
Notes to Chapter 2 370
Notes to Chapter 3 374
Notes to Chapter 4 377
Notes to Chapter 5 379
Notes to Chapter 6 381
Notes to Chapter 7 382
Notes to Chapter 8 384
Bibliography 387

Introduction and Roadmap

In Morals by Agreement, David Gauthier tells what I take to be the most convincing story yet about what it is we might be doing when we are thinking, talking, and behaving morally. [1] Gauthier has a story to tell too about how to think, talk, and behave morally. He urges us to become what he calls 'constrained maximizers'. [2] He has another story to tell about how to 'bargain' - he recommends 'minimax relative concession' [3]; and yet another about how to agree on what each of us may be allowed to bring to the bargaining table in the first place - he puts his own twist on the Lockean Proviso. [4] (Thus by 'bargaining' he means what most of the rest of us mean by resolving issues of distributive justice.) But these latter stories will be of only marginal interest to us here. Here our interests will be largely meta-ethical, rather than meta-economic.

I begin, then, with a largely supportive rehearsal of Gauthier's meta-ethical story, including its domain of discourse, its motivation, its methodology, its conception of rationality and, finally, its pivotal move. What shall be of especial interest, however, is less the substantive details of Gauthier's account than its explanatory schema, the central feature of which being that there are certain 'games' - games paradigmatic of the human interpersonal condition - the winning strategy for which involves the performance of an operation on some feature of one's own psychology. Of these the most interesting will prove to be what might be called 'second-order' self-modifications, that is, operations performed on one's own psychology designed to ensure the development of - and/or once developed, the protection of - some of these aforementioned first-order modifications.

But of these, in turn, I will be focussing on second-order self-modifications designed solely to ensure the unreadoptability of some feature of the original psychology. I call these features - that is, those it behooves us to render unrecoverable - 'self-effacing'. And my claim is that a comprehensive analysis of the rationality of self-effacement can make good on a promise of incalculable explanatory power.

In the course of this analysis, however, it will emerge that there is a paradox within the rationality of self-effacement, the resolution of which may require a modest revision, but a revision nonetheless, to the conception of rationality with which we started out. This will thrust us squarely into the longstanding debate - only now armed with an agenda of our own - over whether prudence is to be characterized strictly in terms of current preferences which are simply higher-ordered, or rather in terms of first-order preferences which can also be anticipated and/or remembered.

I conclude (and also intersperse) with a series of comments on what ramifications, if any, our findings might have for the completion of the meta-ethical program with which we (and Gauthier) started out, for the debate over the conditions on moral considerability, for naturalized epistemology - and, therein, for the residual problem of skepticism - and, finally, for the possibility of yet another convincing story about what it is we might be doing when we are thinking, talking and doing philosophy itself.

But for all this talk of Gauthier, I want this to be understood less as an exercise in game-theoretic thinking about morality than as an enquiry into the logic of self-effacement. Accordingly I have elected to preface my enquiry with what I hope will serve as both a narrative hook by which the reader may be drawn in in the first place and, once there, a constant reminder of what it is about the human condition that this enquiry is about. I call this hook - The Man in the Glass Booth.

1. THE MAN IN THE GLASS BOOTH

He is surrounded by floor-to-ceiling panes of reinforced, bullet-proof glass; and this glass booth, in turn, by heavily armed Sephardic guards. Without these he would be torn apart by the gallery.

He stands proud and erect, recounting in bone-chilling detail, and with great relish, the indignities that he, Adolf Karl Dorff, once had the privilege of visiting upon the members of their families. Some of them he even remembers by name.

The circumstances of his capture have already been established by the Israeli prosecutor. With the Allies at the gates, his day of reckoning at hand, it is assumed he assumed the garb, and the identity, of one of his victims, one Arthur Goldman; and was so successful in this subterfuge he was able to re-establish himself, in time, as one of New York City's most successful real estate developers. In fact so successful was he in this subterfuge - indeed it was probably a condition of that success - that he became the Jew whose identity he assumed; became him even in his own mind. So that even twenty years later, Dorff (now Goldman) continued to live in abject terror that Dorff, the monster who once persecuted him - and in all the intervening years was never found! - was lying in wait for him still, somewhere in the streets of Manhattan.

But the men who were lying in wait to kidnap him were not the agents of this imagined nemesis, but of Mossad. And, in fact, it was only after months of intensive 'deprogramming' by Israeli psychologists that the wolf was finally coaxed from his sheep's clothing.

Having shed this clothing, he has mustered no defense. For ridding the planet of 'vermin' he thinks he needs none. Still, notwithstanding his proud confession, justice must satisfy itself it has the right man. It was with considerable difficulty but, alas, the dental records of both Goldman and Dorff are finally in the hands of the court. With those of the accused they are compared, with the following results:

The man that the man now in the glass booth once was is not Dorff, the Nazi. He is Goldman, the Jew!

(The Man in the Glass Booth, by novelist, playwright and consummate actor, Robert Shaw, was published in 1967. In his 1975 American Film Theatre adaptation, Director Arthur Hiller took considerable liberties with the story. The foregoing synopsis takes modest but further liberties still.)

1. Questions:

When I first encountered The Man in the Glass Booth, I was too moved by more primitive emotions to ask any of the following questions:

Did Goldman's gradual identification with Dorff, in the tortured privacy of his mind, take place only after his rescue and emigration to America? This seems most likely, because it is at least familiar. For many of the survivors of the Holocaust - henceforth I will use the more 'politically correct' term Shoah [5] - death had become so routine that their world was literally divided into two camps: those who had been, or shortly would be, killed, and those who were doing the killing. So, in Goldman's already damaged mind, from sharing responsibility with Dorff, to sharing his persona, could have been a very short distance.

Or had the process already begun while he was still a prisoner of the Nazis? Indeed, had it already been completed by the time the camps were liberated? This would not have been any the more horrible. Nothing could be. But it would have been more bizarre. How would it have been possible for him to sustain the mental life of his persecutor while all the while enduring the physical privations of one of his own victims?

The two men must have born a likeness - else why would Mossad have suspected him? So was there an actual moment, just before the camp was liberated, at which he imagined he had just donned the garb he had been wearing all along? Was he worried, right up to the time he boarded the boat for America, that one of his surviving victims might recognize him and turn him in?

And how long did it then take to become Goldman again? Was he not suspicious of the ease with which he had made the transition? And if the Israeli 'deprogrammers' were able to resurrect Dorff from within Goldman, why could they not resurrect Goldman from within the resurrected Dorff? Was it because, having found what they were looking for, they never thought to look further? Or was it because Dorff was never quite extinguished by Goldman, but Goldman was no longer even obtunded within Dorff?

But, as I say, at the time I never thought to ask any of these questions. I suppose they did not seem important. I am not sure even now that they are.

2. Reactions:

In fact my initial reactions were threefold, successive and, in each case, incredulous. The first - cliched though it may be - was that human beings could do such things to one another, and could do them with the mephistophelian delight with which the man in the glass booth recounted them. (This in spite of having been virtually weaned on stories about the Shoah.) The second was that it is even possible for the human mind to perform the kinds of gymnastics which, were the story to be believed, would have to have been the psychological history of the man in the glass booth; but that such gymnastics, be they possible or not, should ever be necessary. And the third was that such gymnastics, be they necessary or not, could ever be defensible.

Let us canvas these reactions for their philosophical grist.

With respect to the first reaction, revisionists are quick, but right, to point out that neither the material 'shortages' nor the psychological 'stresses' that characterized the camps were unique to them. Our horror, then, must lie in what we take to have been the magnitude and/or the systematicity of these privations. And these, of course, are precisely what the revisionists dispute. Besides, it is easy enough to say that human beings ought not prima facie to do such things to one another. But, claim they - and it is hard to disagree - it has always been a live and open question, both in ethics and in social/political philosophy, just what, exactly, is the force of this codicil '_prima facie_'. Still, surely the least that can be got out of our (thankfully widespread) outrage over this particular episode in 'man's inhumanity to man' is that any ethical and/or social/political theory that can be pressed to countenance anything as vile as the Shoah ought, on that account alone, to be at least double-checked.

The second reaction is akin to but more nuanced than the first. Here the question, What exactly do we owe our fellow human beings? is simply being reposed with respect to a more specific utile, namely the material and psychological conditions required to prevent the disintegration of one's personal identity. Let us set aside, for the moment, Goldman's own role, if any, in this disintegration. Surely no one would deny that

1) the man that the man now in the glass booth once was suffered the disintegration of his personal identity; that
2) notwithstanding that this may have been the lesser of two evils, that man was harmed by this disintegration; that

3) this disintegration was caused, in a morally relevant way, by the material shortages and psychological stresses of the camp; and that
4) those responsible, in a morally relevant way, for those privations are therefore likewise responsible, at least in part, for that harm.

So if it is grist we want, this second reaction would, at the very least I think, lend support to the grievance - laid by marxists and feminists against capitalism and patriarchy respectively, and by both against traditional liberalism - that the distinction between positive and negative rights, and the cognate distinction between acts of commission and omission, are indeed naive and inadequate.

But, as before, it is one thing to acknowledge that the psyche, no less than the body, needs room in which to grow and sustain itself. It is quite another to fix upon just how much of this personal 'space' each of us can lay claim to, and at what cost to ourselves we may be required to provide such space to one another. After all, the exigencies of war - and _a fortiori_ of what was, for most intents and purposes, _total_ war - are such that they not only require that people _make_ sacrifices, both physical and psychological, but also that people sometimes be sacrificed, both physically and psychologically. But again, surely the least that can be got out of even this 'nuanced' outrage over the conditions that must have driven Arthur Goldman to fashion himself in the image of Adolf Dorff is that any ethical and/or social political theory that can be pressed to countenance anything as vile as the nazification of a Jew ought, on that account alone, to be regarded as at least suspect.

3. Personal Identity:

We have just seen that, grist though there may be here for reflection on some rather core ethical, social and political issues, there is nothing uniquely informative in this regard to be found in either of these first two reactions. But let us see whether there might be metaphysical grist here instead.

To begin with, let us take it as given that - whether against extant natural law or merely retroactive positive law - the deeds recounted by the man in the glass booth were crimes. [6] One question one might have, then, is this. Was it the man now in the glass booth who committed those crimes? And yet another is, if so, does he remain accountable for them? Of course one might hold that establishing that the perpetrator and the accused are one and the same person need be neither necessary nor sufficient for the latter's being accountable for the deeds of the former. Or even if necessary, certainly not sufficient. But, for the time being at least, let us ignore such subtleties. Let us just say that if the man in the glass booth is Dorff rather than Goldman then he is - and it is *he* who is - accountable for the most heinous crimes against man, woman, child, and God.

So, if we hold a person just *is* his body, then the dental records would have settled the issue. The man in the glass booth is Goldman, not Dorff, and that is all there is to it! If on the other hand we think a person is his psychology, then the man in the glass booth *might* be Dorff, depending on, of course, to which of the myriad versions of the Psychological Criterion View we would like to subscribe. To wit:

It has been almost exactly three hundred years since John Locke first proposed that, though physical continuance may serve as *evidence* for psychological continuance, it is psychological unity and continuance - and more particularly, thought he, the unity of apperception and the veridicality of memory - which are the *criteria* of, respectively, synchronic and diachronic personal identity. [7] How so? Because, Locke insisted, "Person is a forensic term," by which he meant, "[pertaining to the] appropriating [of] actions and their merit." [8]

That is, it may be that souls are something distinct from persons. (And, for all we need care, maybe selves are even some third thing, distinct from both.) It may be souls are not mere bare haecceities, they are distinguishable, albeit only by God, and notwithstanding they are uninformed by their worldly sojourn God - who is surely entitled to meet out Her Own justice - rewards and punishes them in the Here-after as only She sees fit. It may even be souls are informed by their worldly sojourn, that the mystery of soul/body interaction is a mystery only to us, and therefore God's justice need not be all that alien to us. But even if any or all of this were true, none of it would relieve us of our responsibility to pass our judgments, both on others and on ourselves. And it is for these purposes that the notion of a person - as distinct from a soul (and/or perhaps a self) - is designed and, therefore, rightly reserved. Or, as Locke himself put it:

> It is not therefore unity of substance that comprehends all sorts of identity, or will determine it in every case; but to conceive and judge of it aright, we must consider what idea the word it is applied to stands for ... which, if it

> had been a little more carefully attended to, would possibly have prevented a great deal of that confusion which often occurs about this matter, with no small seeming difficulties, especially concerning personal identity. [9]

So rather than allow our (admittedly rampant) *mis*uses of the word 'person' to contaminate our judgments about moral accountability, we should instead allow our (by and large *reliable*) judgments about moral accountability to discipline these abuses. Locke's prince and cobbler [10], John Perry's Julia North [11], Bernard Williams' mad scientist [12], Derek Parfit's teletransportation scenarios [13] - these are all designed to help us distinguish the *truly* criterial from the merely (and contingently) evidential. And, claim Locke, Perry, Parfit *et al*, what this puzzle-case methodology reveals to us is that it is quite rightly by psychologies rather than bodies that we track moral accountabilities. So, claim they, it is by psychological identity rather than bodily identity we ought likewise to track personal identity.

But, as Bishop Butler was quick to point out [14] - and as Thomas Reid demonstrated even more convincingly [15] - Locke's 'Naive' Memory Criterion View is either self-defeating or else hopelessly circular. Why? Because the only way to distinguish between veridical memories and pseudo ones is with reference to either *physical* identity - e.g. It was *his* eyes that witnessed the event he now remembers! - or else personal identity - e.g. It was he to whom the event he now remembers originally happened! Thus was set in motion a series of iterative refinements to the Lockean view - e.g. by Anthony Quinton [16], by H.P. Grice [17], by John Perry [18] - one of the latest, though undoubtedly not the last, being Derek Parfit's in Reasons and Persons. [19]

Intent upon circumventing the circularity problem once and for all, Parfit proposed a bare psychological similitude relation called 'Q'. (What we are to imagine, presumably, is that, baldly put, psychological states can be projected onto a screen, compared, and then assigned a similitude point count.) He then urged that person-stage C be deemed the legitimate continuer of person-stage A just in case:

1) person-stage C is (what he calls) Q-memory related to person-stage A [20] - and sufficiently so related to make the identity claim at least plausible [21] - and
2) there exists neither
    a) some person-stage D, co-temporal with person-stage C, who is either
        i) more Q-memory related to person-stage A than is person-stage C, or

ii) as Q-memory related to person-stage A as is person-stage C, nor

b) some person-stage B, co-temporal with person-stage A, who is either

i) more Q-memory related to person-stage C than is person-stage A, or

ii) as Q-memory related to person-stage C as is person-stage A. [22]

Q-memory relatedness, in turn, is defined in terms of a) psychological similitude and b) causal connexity. But

a) the conditions on psychological similitude, in turn, make no reference whatever to the _real_ (i.e. mind-independent) referents of the mental states being compared; and

b) the causal connexity condition requires only that this psychological similitude be in _some_ way - though, Parfit allows, not just in _any_ way [23] - causally occasioned by the psychology of person-stage A.

For Parfit, then, personal identity claims require the satisfaction of 1) the Psychological Similitude

Condition, 2) the Exclusivity Condition (or Non-Branching Proviso), and 3) the Causal Connexity Condition. [24]

Parfit's is, as I say, one of the latest words on personal identity. [25] But a decade earlier David Lewis proffered the counter-suggestion that we can as readily - and perhaps even better - accommodate the fission and fusion cases that have inspired the Exclusivity Condition by acknowledging instead that there are, or were and remain, two persons there all along! That, in other words, any number of persons be allowed to share a common person-stage. And that, in other words still, we jettison the Non-Branching Proviso and content ourselves instead with only 1) Psychological Similitude and 3) Causal Connexity. [26]

Lest some stone remain unturned, however - and following the lead of Eli Hirsch [27], Daniel Kolak and Ray Martin [28] - I have been urging instead the sufficiency of 1) Psychological Similitude and 2) Exclusivity. [29] Let us confine ourselves to these three options, i.e. Parfit's, Lewis' and my own; and let us see how, with respect to the case at hand, each would cash itself out. With respect to whether

1) Goldman-turned-Dorff (G-D) - i.e. the prisoner fancying himself his own persecutor - can be the legitimate continuer of Goldman, or whether
2) Goldman-turned-Dorff-turned-Goldman-again (G-D-G) - i.e. the persecutor masquerading as one of his own victims to avoid retribution - can be the legitimate continuer of G-D, or whether
3) Goldman-turned-Dorff-turned-Goldman-turned-Dorff-again (G-D-G-D) - i.e. the wolf having shed his sheep's clothing - can be the legitimate continuer of G-D-G,

Parfit, Lewis and I are of a mind. In the absence of any further information about the incrementality of these various metamorphoses, in none of these cases is the 'sufficiency' condition met. So *prima facie* at least, the man in the glass booth cannot be Dorff.

That said, all three of us allow for identity-claims across temporal breaches. So, just as

4) G-D-G can be the legitimate continuer of the original Goldman,
5) so too can G-D-G-D be the legitimate continuer of G-D.

Of course we are also of a mind that personal identity is transitive. (Could any identity relation fail to be?!) So if

2) G-D turned G-D-G again in increments, each of which satisfied the sufficiency condition, and if
3) G-D-G turned G-D-G-D again in likewise sufficiency-satisfying increments,

then, provided that, as it happens, neither

a) the Exclusivity Condition nor
b) the Causal Connexity Condition

are violated - or just not the latter for Lewis or just not the former for me -

5) G-D-G-D *would* be the legitimate continuer of G-D.

So if G-D could be the legitimate continuer of Dorff, then so could G-D-G-D. But then, by parity of reasoning, if

1) Goldman turned G-D in sufficiency-satisfying increments -

and if, as it happens, neither

a) Exclusivity (for me) nor
b) Connexity (for Lewis) nor
c) both (for Parfit)

were compromised - then, notwithstanding the apparent violation of the sufficiency condition, G-D would be the legitimate continuer of Dorff. Similarly, if

3) G-D-G turned G-D-G-D, again in sufficiency-satisfying increments -

and if, as it happens, neither

a) Exclusivity (for me) nor
b) Connexity (for Lewis) nor
c) both (for Parfit)

were compromised - then, notwithstanding the apparent violation of the sufficiency condition,

3) G-D-G-D would be the legitimate continuer of G-D-G.

So, it appears, the man in the glass booth could be Dorff after all!

Nor is there anything peculiar in any of this. Most of us are the same persons we were as infants, notwithstanding few of us can remember anything of our infancy. And each of us upon recovery from amnesia remains the same person she was prior to suffering it.

But let us suppose at the precise moment Goldman achieved psychological identity with Dorff, the latter had been killed, say by an Allied shell. Then, by all three of our lights - i.e. Parfit's, Lewis' and my own - he would in fact have survived, i.e. in the person of G-D. If on the other hand he had not been killed, then on Parfit's and my view - but not so on Lewis' - he would not have survived. This is because, given our Non-Branching Proviso, we say that there being two of them there are none of them. Parfit might be able to avoid this result by noting that the man physically continuous with Dorff can claim exclusive clos-est continuer status, in virtue of his stronger, i.e. less attenuated, causal connection - a defense, note, unavailable to me. For me, by contrast, it would have to be that, even if the man psychologically _and_ physically continuous with Dorff had been found after all these years, he could not now be held accountable for anything Dorff might have done prior to Goldman achieving co-temporal psychological identity with him since, at that point, Dorff ceased to exist. Of course by the same token neither could _Goldman_ be held accountable for anything Dorff might have done prior to Goldman achieving co-temporal psychological identity with Dorff. But, I concede, the relief of the Israeli Justices at being able to release Goldman would be small consolation for their frustration in being unable to convict the man now found to be psychologically _and_ physically continuous with Dorff.

As we have just seen, Parfit has his own bullets to bite. But if both Parfit's account and my own diminish accountabilities below our need to impose them, Lewis' multiplies them beyond our capacity to bear them. For, according to Lewis, if the man both psychologically and physically continuous with Dorff were to undergo a moral epiphany, burst Jean Valjean-like into the courtroom, and offer to take Goldman's place in the dock, then, rather than being able to release the man currently in the glass booth, the court would be required to try, convict and sentence both men!

What Parfit and I are unclear about, however, is what to say about the following case. Suppose Dorff had not been killed by the Allied shell, but had merely suffered a mild concussion followed by temporary amnesia. Then - ex hypothesi, quite straightforwardly and, as we have seen - Dorff would have survived, i.e. in the person of G-D. But at the precise moment erstwhile-Dorff recovered his memory, would G-D cease to be Dorff? And would erstwhile-Dorff simultaneously fail to recover his own continuance? If so, could erstwhile-Dorff ever become Dorff again? And if so, could he do so in virtue of nothing more than G-D either dying or turning Goldman again?

According to Bernard Williams, for one, it is at this point that it is our teeth rather than the bullet that starts to crumble. For whatever the necessary and sufficient conditions on personal continuance might be, he cautions, they had better not be such that they can be met or defeated by any such highly contingent and trivial fact of the matter, i.e. one which is metaphysically remote and/or epistemically inaccessible to all parties concerned save perhaps God. [30] For the object of our exercise, recall, is to settle on a theory of personal identity which is amenable to our real-life, interpersonal, social, political and legal forensic needs. So perhaps - following the lead of many jurisdictions - we should just impose a 'statute of limitations' on how long after having gone missing (amnesia) and been presumed dead (either by duplication or by being more closely continued elsewhere) one will be allowed to return from the grave, so to speak, with her forensic entitlements and liabilities intact.

Are moves like this altogether too ad hoc? Perhaps. But it is doubtful there will ever be a perfect solution to puzzle cases like these. Nor should we expect there to be. But, while conceding that "the method of science fiction has its uses in philosophy," ought we to wonder, along with W.V.O. Quine, "whether the limits of the method are [being] properly heeded [here]"? Is Quine right to suggest - as he does in a review of an anthology of just such puzzle case musings - that

> [t]o seek what is 'logically required' for sameness of person in unprecedented circumstances is to suggest that words have some logical force beyond what our past needs have invested them with. [31]

On all these options have countless philosophers laboured long and hard. But, alas, our response to The Man in the Glass Booth relieves none of their burden. As Williams has pointed out, our intuitions on these scores are, at the same time, both recalcitrant and unstable.

[32] So unless our response to The Man in the Glass Booth offers something by which to loosen or stabilize them, we have plenty and enough to put through the personal identity debate mill as it is.

But long and hard too have laboured philosophers to settle the alternative issue. Suppose, being unable to reach consensus on just who, exactly, the man in the glass booth might be, we have nonetheless elected to treat that question, and the question of accountability, as entirely separate and distinct. So, we might ask, notwithstanding the man now in the glass booth may not be the person who committed these crimes, might he be held accountable for them nonetheless? After all, he thinks he committed them; he is as proud of having done so as he would have been had he committed them; and so he is certainly as likely to 'repeat' them as would the man who actually committed them in the first place.

Nor need it be only those with strictly consequentialist concerns who could be tempted by such a philistine end-run around the question of personal identity. For even those with purely retributivist intuitions might hold that what is punishable is not so much the particular mind that as a matter of historical fact gave rise to the crime, but rather a) the mind that would have given rise to the crime, coupled with b) the facticity of the crime itself, notwithstanding that c) the particular mind referred to in (a) may not, as it happens, have given rise to the particular crime referred to in (b). But, alas, our retributivist and consequentialist intuitions are probably no less recalcitrant, and no more stable, than those pertaining to personal identity. Does our response to The Man in the Glass Booth offer to unsettle or stabilize them? Probably not. Still, let us keep these questions hovering somewhere in the periphery of our minds. Who knows when we will encounter something that niggles at us as having relevance?! [33]

4. Crimes Against the Self:

But what remains, recall, is my third reaction.

Above and beyond the crimes against humanity (as standardly understood) that may have transpired here, and our search for someone to hold accountable for them, is there not another crime that has been committed here? Something obscene and profoundly unnatural - no less so than the patricide of Oedipus, the regicide of MacBeth, or the deicide of the Passion - against which all of nature rebels? Something akin to, but not quite, suicide. Perhaps we should call it 'autocide'. But even that does not capture it.

Is it not just this? That we are casting about for someone with whom to profoundly sympathize, and for someone else upon whom to visit unstinted vengeance; and we are saddened and angered, respectively, at being unable to find either? And, what is worse, that neither can the victim find either the perpetrator or himself? And is not the obscenity just this? That this is itself the crime?!

One is reminded of the twisted logic - but the logic nonetheless - of the murderer who reasons he could not possibly have deprived his victim of his future prospects since, being about to be murdered, he had no future prospects of which to be deprived. One is reminded as well of the story of two lovers cursed by a jealous sorcerer, she to turn into a hawk by day, he into a wolf by night. [34] Only in our story it is the sorcerer himself, so consumed by self-loathing, that by day and by night respectively he becomes what by night and by day respectively he would kill. And so across an infinity of dusks and dawns he pursues himself, catching a glimpse of his quarry for one fleeting moment just before he becomes it.

Such, or something similar, are the wages of sin. Or so we are told. So maybe self-betrayal - if such this be - is a sin. Plato taught as much. So does Christianity. The soul languishes for all eternity in unrequited yearning for reunion with the betrayed self.

But human beings have no souls. Or so we have told ourselves. So why should we be bothered by any such metaphysical nonsense?! Some lives go better than others. From '33 to '45 Nazis fared better than Jews. When the camps were liberated the advantage shifted. So given a choice why not be a Nazi until '45 and then a Jew thereafter? Most of us are not given the option. Goldman and Goldman-turned-Dorff were. They seized it. Are we not confusing outrage with envy?

This is one intuition. Here is another. Robert Nozick once offered us access to a marvelous machine. Plug yourself in and you are guaranteed a life which is, technically, at least as good as any other - which means, effectively, the best of all possible lives. [35] Some of us reached for the chord. Most of us did not. Is this just one of those recalcitrant differences? I think not. I think there is an abiding truth in the refrain, "I don't want it 'less I find it myself!" [36] The intuition weakens if what we are offered are greater powers in the quest. [37] And it weakens still further if what we can have for the asking is a world in which there is more to find. [38] But in the same way, and for the same reason, as in a friendly game of poker the rest of us roll our eyes when a novice calls "Everything wild!", neither is life itself meaningful without the constraints of our own historicity, our own facticity, our own givenness. Without, that is, there being something that holds firm enough for us to set our shoulder against it, to feel it give, or to feel it not give, as the case may be ...

"Pretty thoughts indeed!" scoffs the other voice within us. "Tell them to the victims of torture, to the victims of pestilence, of famine, of drought. Tell them to the mother watching helplessly as her children wither silently like fallen fruit under the Saharan sun. Console her, if you would, with her 'historicity', her 'facticity', the 'givenness' of her situation! Console Goldman!"

Indeed, argue some feminists, such are the fruits of masculinist philosophizing!

Life is not something to be chosen or forgone. Life is what grounds the very possibility of choosing. Nature has taught us to pay dearly for just this. It is something we have learned to respect about each other. That is why the walk from death row is made in leg irons. [39]

"But people do die for their principles," one voice reminds us.

"Yes," answers the other, "but one wonders sometimes how often they do so because they have misidentified a mere means as an end. Curious - is it not? - how our principles and standards plummet when we no longer have the wherewithall to sustain them. Fortunate too." [40]

Is it simply, as John Rawls suggests, that presupposed in the incumbency of sustaining (what he calls) the 'principles of justice' is our ability to maintain (what he calls) the 'circumstances' of it? And that the same can be said for principles in general? [41] No doubt. But then surely as incumbent upon us as understanding these principles themselves is understanding the transcendental conditions for their possibility!

5. Defensible Self-Projection:

Still, it seems odd, if not downright offensive, to charge Goldman with opting out of his historicity, of his facticity, of the givenness of his situation. [42] And yet we feel much less reticent - do we not? - about making the same judgment of Dorff. So, it might prove instructive to ask, what is going on here?

Clearly one thing that is going on here is we are conflating our judgments on the moral defensibility of each of the two characters' willingness to become the other with our judgments on the respective moral worths of the two characters themselves. So let us block this conflation. Let us suppose instead it was not Jew-haters but heterosexuals who were most likely to survive the camps; it was not Jews but homosexuals who were most likely to survive liberation; and one can, under dire enough circumstances, alter one's sexual orientation. [43] Assumed too, of course, is the reader's sentiments with respect to Jews and Nazis have no counterpart with respect to gays and straights. Thus sanitized the differential - or so I would venture - disappears.

Of course we can denude the thought experiment further by replacing 'gay' and 'straight' with 'left-handed' and 'right'; and we can denude it further still by replacing 'left-handed' and 'right' with 'innies' and 'outies'. But there is a difference between divesting ourselves of our own agendas and divesting Goldman and G-D of theirs. [44] For if we strip Goldman and G-D of all historicity, facticity, givenness, situatedness, interestedness - call it what you will - then that there can be no differential judgment of their respective behaviours will be true but, alas, trite as well.

This is not to say our judgments should be falling one way or the other; or even that they must fall differentially. I claim no privileged access to the mind-set of a Jew braced for years in the maw of the Shoah, much less to that of a National Socialist in the final hours of the Third Reich. It is to say only that so long as there remains something that differentiates Goldman and G-D historically, situationally - call it what you will - so too remain grounds for judging one self-betrayal, if such it be, more defensible than the other. [45]

We tend, I think, to regard the people of Denmark under the Occupation as having been supererogatory in wearing Stars of David in an attempt - not entirely futile as it happens - to protect their fellow citizens. This is not to say we think we would have behaved as nobly, but only it was a noble way to behave. We tend, I think, to regard as somewhat suberogatory those who attempted to deny their Jewishness when the 'selections' at the Auschwitz of the early '40's were re-enacted at the Entebbe of '75. This is not to say we think we would have behaved more nobly, but only there was a more noble way to behave. And as having got it just about right, I think, is how we tend to think of those who put their all on the line to defend the State of Israel in '48, '56, '67 and '73. [46] The reader may have other intuitions. If so, I will not dispute them. My point is simply this. To deny outright there are appropriate, more appropriate, and inappropriate ways of projecting one's self and one's values into the future, is little more than existential nihilism!

What is crucial to note, however, is that in passing judgment on the man in the glass booth, we have been trying to apply this principle to the limiting case. Indeed, that is what makes the case so intriguing for us. But from the acknowledgement that we pass judgment in this case only with a presumption bordering on obscenity, does it follow we should refrain altogether?

I think not. I think if we can decide - and I think we have got to decide - in this, the most penumbral of cases, whether Goldman can be allowed to assume the persona of Dorff, notwithstanding the physical and psychological duress under which he did so ...

... if we can decide - and I think we have got to decide - in this, the most penumbral of cases, whether Goldman-turned-Dorff can be allowed to escape responsibility for his deeds, imagined or otherwise, by becoming Goldman again, notwithstanding the unwitting rectification affected by this second becoming ...

... then, I think, we should be well on our way to an account of what are and are not defensible ways of projecting our selves and our values in what is, thank God, the less taxing core of our self-projecting experience. [47]

So the question, in a nutshell, is this. If we are to allow people like Dorff to escape responsibility for their actions by means of personal self-effacement, then we must likewise allow people like Goldman to take on responsibility for the actions of others by means of personal self-effacement. So on the one hand we are tempted to deny re-sponsibility to both of them. On the other hand drugs are nigh on the market that efface the memory of deeds done for some period of time after the ingestion of the drug, including the ingestion itself. Can we allow the mens rea condition to be defeated so readily? [48] Might we need to make a social policy decision to discourage such ploys, a decision which, notwithstanding its violence to our metaphysics of persons, might nonetheless have to trump? Or would it be better to simply alter our metaphysics of persons?

This, then, is what I take to be the philosophical grist offered by The Man in the Glass Booth. But with what shall we mill it?

6. Gauthier:

As I said at the outset, David Gauthier tells what I take to be the most convincing story yet about what it is we might be doing when we are thinking, talking, and behaving morally. I will have a great deal to say about this story in due course. For now, however, suffice it to say that, for Gauthier at least, answering the question, Why be moral? is best served by posing what he takes to be the logically prior question, Why become moral? That is, in the tradition of Glaucon and Hobbes before him, Gauthier asks us to imagine ourselves as we were - or at least would have been had "there [ever been] such a time" [49] - in our pre-moral condition, and to then ask whether it would have been defensible - which he takes to mean 'rational' - to become moral. Thus the answer to the question, Why be the way we are? is that, given the way we once were - or would have been were we not the way we are - it was (or would have been) rational to become the way we are.

But, it is to be noted, there is a difference between asking, Why be a certain way? and, Why become that way? For it might be that, from the perspective of someone who already is that way, the first question makes little if any sense. "'Why am I moral?', you ask?

Because, dammit, that's just the kind of person I am!" That is, *being* a certain way - in this case being moral - may feel as much an existential choice as deciding to *become* moral. Nor is the question, Why *be* moral? reducible to the question, Why *remain* moral? For remaining is just a special case of *becoming*. It is a decision to *continue* a certain becoming.

So here are three questions which may or may not prove embarrassing for Gauthier. We will have to see. First, if my having *become* moral and my having *been* moral all along are phenomenologically indistinguishable from one another, why should I believe the one, i.e. that I only *became* this way, rather than the other, i.e. that I had always *been* like this? Let us call this the 'underdetermination problem'.

Second, if my being moral cathects an existential choice to *remain* moral - and does so *irrespective* of whether I became moral or had been moral all along - then, apart from armchair curiosity, of what possible interest could Gauthier's story be to me *qua* moral agent? Let us call this the 'relevance (or prescriptivity) problem'.

And third, if Gauthier's story *is* relevant, then it can be so only in virtue of there being some kind of perspectival *privilege* to my putative pre-moral condition. In which case, in virtue of what is this pre-moral condition privileged? Let us call this the 'privileged perspective problem'.

What is important to note, however, is that with regard to the rationality of becoming there should not be anything peculiar about morality _per se_. Gauthier's analysis - and these three problems therewith - should be as readily applicable, _mutatis mutandis_, to the defensibility of Goldman's becoming G-D and to G-D's becoming G-D-G again. That is, if, as it turns out, it could only be rational for me to remain moral provided it would have been rational for me to become moral had I not been moral already, then can we likewise say that it can only be rational for G-D to remain G-D provided it would have been rational for Goldman to become G-D? And, _mutatis mutatis_, that it could only be rational for G-D-G to remain G-D-G provided it would have been rational for G-D to become G-D-G? Obviously not! But why not?

Suppose that, as it happens, it was irrational for Goldman to become G-D - because, let us suppose, as Goldman he in fact preferred there be no one at all than there should be yet another monster the likes of Dorff. That is, he simply made a mistake. Of course other than from the testimony of some third party, would there be any way G-D could become convinced of his own true etiology? Probably not. But let us suppose this underdetermination could be resolved. Let us suppose, on grounds independent of the phenomenology of being G-D, G-D becomes convinced he was once Goldman, just as, on grounds independent of the phenomenology of being moral, I have become convinced I was once the selfish 'bastard' described to me by Hobbes, and that life thereas was, as it was for Goldman, "solitary, poor, nasty, brutish and short". [50] Even if G-D became convinced it was a mistake for Goldman to have become G-D, would it be rational on those grounds alone for G-D to become G-D-G again? We have already observed that, with the Allies at the gates, it might have been - indeed it probably was - rational for G-D to become G-D-G again. But this was on independent grounds, i.e. independent of its having been irrational for Goldman to have become G-D in the first place. So, it would seem, the irrationality of Goldman's becoming G-D - if indeed it was irrational - does not penetrate to inform the rationality of G-D's remaining G-D.

Similarly, suppose it had been indefensible for G-D to have become G-D-G - because, let us say, as G-D he should have preferred there be no one at all than there should be yet another 'vermin' the likes of Goldman - and that G-D-G becomes convinced, on likewise phenomenologically independent grounds, of both this etiology and its indefensibility. Would it have been rational on those grounds alone for G-D-G to become G-D-G-D again? Obviously not. But then if, as we have just seen, it can be rational not to repair an irrationality - if, that is, the rationality of repairing to some prior state is determined on the basis of what is rational for me now, and quite independent of what may or may not have been rational for me then - then to what purpose is Gauthier urging us to imagine ourselves in our pre-moral condition?

Clearly he must assume - quite rightly, I think - that there is enough of our pre-moral condition still extant within us that we remain, to some degree at least, free to evaluate our erstwhile metamorphoses into moral beings, and to regulate those metamorphoses accordingly. What The Man in the Glass Booth tells us, however, is there are situations of existential choice in which the phenomenology of the existential choice that gave rise to it is nowhere to be found in it, save perhaps as a story, independently grounded, about the etiology of that existential choice. How, if at all, is such information to be accommodated?

Marxists and feminists have been reminding us that our values are - if not wholly then at least in part -

socially constructed. No doubt they are right. But what are we to make of this? Deconstructing our values and assigning them their social etiologies may be useful to social engineers. But what is the individual valuer to do with this information? Richard Rorty, among others, tells us that recognizing the radical contingency of our values releases their grip on us, thereby multiplying our options. [51] No doubt that is true too. How else does raising our consciousness about, for example, exploitative advertising techniques, make us more intelligent consumers? But from what - other than other, as yet undeconstructed, values - are we to select from among these new-found options?

This is an old saw, but not, on that account, any the less troubling. Nor is it less troubling for the sociobiologist, for example, than for the marxist or feminist. Each has a story to tell. And there is probably truth in all of them. I think the story I will be telling is more comprehensive. But with respect to these problems of underdetermination, prescriptivity, and perspectival privilege - we are all of a piece.

Rightly or wrongly - probably wrongly - here is a way to understand Kant. Recognizing, along with Hume, the radical contingency and inherent instability of who we are - which is just to say, the bundle of substantive values constitutive and definitive of ourselves - and yet not at all content to leave the matter at that, Kant observed, as we have ourselves just done, that existential choice requires at the very least _formal_ constraints on permissible self-projection. And these formal constraints, he thought, are nothing more nor less than those of practical reason. Thus whereas for Gauthier, as we will see, rationality is simply that feature of the mind in virtue of which it more or less reliably takes itself from whatever it wants to the means to achieving that 'whatever', for Kant - and, as we have just seen, he seems to be right - such instrumental rationality is inadequate even on its own terms. For instrumental rationality, by its very definition, can provide no critical perspective from which G-D might evaluate his having made the transition, nor from which G-D-G might evaluate his having become _him_self, nor from which, in my post-moral condition, I can evaluate my having become _my_self. What is required, thought Kant, is a perspective which is neither Goldman-esque nor Dorff-ish, neither pre-moral nor post-moral, but rather indifferent with respect to any _substantive_ characteristics, and concerned only with the purely formal constraints on allowable relations _between_ such substantive characteristics. That Kant got those constraints right, I sincerely doubt. But that this, or something akin to it, was what he was after is, I think, a not unreasonable reading of an otherwise rather curious project.

In any event, let us make it our project.

2. THE CONTRACTARIAN SCHEMA

1. The Schema:

There is a tradition in meta-ethics - a tradition as old as the Glaucon of the Republic [1] - according to which our moral dispositions can best be understood as

> an otherwise unremarkable subset of a set of rational, strategic, intramental responses to corresponding patterns of otherwise dilemmatic, game-theoretic interactivity,

patterns thought to be, if not endemic to then at least paradigmatic of, the ecological condition of any number of organisms or systems, but most especially human beings. Thus by morality being 'unremarkable' is meant only that, once we have identified in virtue of what morality is a subset of this set, and once everything (that can be) has been said about the set, there is little, if anything, about the subset that remains to be said.

2. Inter-Level Reductionism - Objections and Replies:

We can begin to decipher this mouthful by noting that by this being a 'meta-ethical' story, as distinct from an ethical one, we mean it addresses itself not to any question that might be posed in the object-language, i.e. in the language of morals itself, but rather to what it is we might be *doing* when we are thinking, talking, and behaving morally - when we are, as it were, 'oughting'. In answering this question, then, we must, on pain of explanatory circularity, do so without recourse to the language of morals itself.

Let us call this project 'inter-level reduction'. Just why, or even *that*, anyone should be interested in such a project, remains to be established.

It will prove vitally important, however, to have distinguished *this* constraint on our enterprise and one altogether different were the object of our exercise 'int*ra*-level reduction'. That is, it is one thing to claim - rightly or wrongly - that, for example, when a young man, sitting and chatting with his lover on a bus, absently trails a finger along the bridge of her wrist, he is actually acting out an nth-ordered mating ritual designed to reinforce an (n-1)-ordered move in the competition for reproduction. (*Mutatis mutatis*, this is precisely what we *are* saying about morality and maximizing on the satisfaction of preferences.) But it is quite another to claim that, when one professes his love this way, what he means, and all he means, is, "I'm attempting to elicit your reproductive cooperation!", or anything of the sort. (*Mutatis mutandis*, this is precisely what we are *denying* about morality and maximizing on the satisfaction of preferences.) What is more, we hope to show *why* intra-level reductionism is false. [2]

Let us suppose we succeed. Should the non-reductionist take any delight in our success? I think not. What motivates the non-reductionist, I think, is a passionate commitment to the defense of the dignity to which she thinks we are entitled in virtue of our being moral beings *categorically*, a dignity to which, on her view, we would *not* be entitled when we are making heartfelt moral judgments were we the grasping, self-absorbed insects represented to us by Hobbes, i.e. were we merely performing some kind of egoistic utility calculus. But, we would remind her, this being 'entitled' or 'not entitled' is itself a moral judgment. So if, *ex hypothesi*, we succeed in our efforts at inter-level reduction, likewise will we have succeeded in explaining away why she is inclined to feel as she does about the value of human dignity, and about what it is that provides that dignity.

That said, the following problem immediately presents itself. If for every intuition that opposes our theory our theory has an explanation for why we have that intuition, and if those intuitions are therefore to be disallowed, then what, if anything, would be allowed to count as evidence our theory is mistaken?

Here is another way of raising the same difficulty. Trite but true, ethicists are used to doing ethics. And by ethical reflection what we mean is, roughly:

> the attempt, by means of the widest possible reflective equilibrium, to induce from the core of our moral judgments the rule, principle, or algorithm underlying those judgments so as to position ourselves to extrapolate that rule, principle, or algorithm into the moral penumbra. [3]

But in what we have been calling 'meta-ethical' reflection, those very judgments are, _qua_ inputs to any _explanans_, of a piece suspect, since they are precisely among what make up the _explananda_. To be sure, grants our detractor, gains in comprehensiveness are laudible; but, she adds, surely not at the expense of precluding falsifiability!

That our theory fails the test of falsifiability would be an awkwardness for us, to say the least, if, but only if, it purported to be a set of historical claims. It does not. In fact this disavowal is precisely the intent of the codicil, "can best be understood as". No doubt on occasion we will lapse into the language of historicity. We will use words like 'genesis' and 'etiology', and phrases like 'natural selective pressure', many of which both connote and denote that we are making claims about how, as a matter of historical fact, human beings actually came to have the moral dispositions they do. Of course we could say that all such references are inadvertent. But this would amount to little more than a confession of editorial sloth. The fact is all of us believe - Hobbes, Gauthier, myself - that the 'responses' referred to in the above schema actually happened. So what is being offered is an account to the best explanation. It is just that we do not want to hitch our wagon to this historical hypothesis, nor to overly preoccupy ourselves with a quest for such data as might confirm or disconfirm it.

For our central claim - Hobbes', Gauthier's and my own - is that even if the world had come into being only yesterday - and human beings along with it with their moral dispositions already in place - our schema would explain why, on purely non-moral grounds, they would be well-advised not to abandon those dispositions. And/or if they happened not to already have them, our schema would explain why, on purely non-moral grounds, they would be well-advised to acquire them with all possible dispatch.

However the key 'if' here is the latter one, i.e. the counterfactual. If we happened not to have the moral dispositions we do, our schema would explain why we would be well-advised to acquire them. But, one might well ask, given that we have in fact already acquired them, of what possible relevance could this counterfactual be to us?

Our intuition, I suppose, is this. Recall the story of the two siblings who are told they can each have one of the two pieces of leftover cake in the refrigerator. The elder takes the bigger one for himself, to which the younger cries out, "Hey, that's not fair!"

"Why?" asks the elder slyly, "What would you've done?"

"I'd have taken the smaller one," answers the younger guilelessly, "and left the bigger one for you."

"Well, then," announces the elder triumphantly, "what are you complaining about? That's exactly what you've got!"

Of course if the younger sibling had already accepted the fairness of his lot - as, mutatis mutandis, our already-moral selves have done - he would not have protested in the first place. And by and large neither do we. A condition of his protest, and our enquiry, being meaningful, then, is there being enough of our pre-moral selves still extant within us (and him) that it would occur to us (and him) to ask the question in the first place. In other words, were we content with our moral dispositions as we find them, the entire Gauthierian story would be more than a tad redundant. But we are not content. And there is the rub. Our ethical intuitions are veritably riddled with tensions, most particularly between our selfish and other-regarding concerns, but more generally between our consequentialist concerns and our non-consequentialist ones. Consequentialists claim our non-consequentialist intuitions can be accommodated by their theory; that they have been instrumentally adopted - and rightly so, say they - in the service of purely consequentialist ones. Kantians concede that many of our consequentialist concerns cannot be accommodated by their theory; but, they insist, these can be dismissed as simply falling outside the domain of moral discourse. What is sorely needed, I think, is an overarching theory, one that can not so much adjudicate between these claims as make sense of them both. And Gauthier's account, as we will see, promises precisely that.

That said, there remain at least a dozen key terms in this schema that have yet to be cashed out. For example, just what exactly is included under the rubric 'moral dispositions'? And, perhaps more significantly, what is excluded? What is meant here by rationality, by strategy, by intramentality, correspondence and patterns? What is intended by dilemma, interactivity, games, theory, organisms, systems, and ecological conditions?

But, just as there is a reason why "Show and Tell!" is not called "Tell and Show!", let us ostend first and theorize later.

3. The Prisoners' Dilemma:

Of these 'paradigmatic' games aforementioned in the above schema, probably the most exhaustively reviewed, in terms of the development of the tradition, has been the Prisoners' Dilemma (or PD). [4]

Two partners-in-crime, Column and Row, are being held in separate interrogation rooms. This isolation is meant to ensure their actions are 'independent'. As we will see, just exactly what this involves will prove a matter of considerable dispute. [5] For now, however, let us just say that, whatever it involves, the wall separating the prisoners will do the trick.

Each is told that if neither rats on the other they will each serve two years. If both rat, they will each serve three. But if one rats while the other keeps the faith, then the ratter will walk while the rattee, i.e. the one ratted on, will serve a full five.

There being no honour among thieves - this to satisfy what we call the 'non-tuism' condition [6] - Column reasons, quite rightly, that regardless of what Row does it is best for Column that Column rat. But, mutatis mutandis, similarly reasons Row. Thus, as a result of the combination of their respective 'straightforward maximizing' (SM) rationalities, they both end up with the penultimately worst outcome; whereas if they had not been so straightforwardly maximizing, they might have ended up with the penultimately best.

Nor would it help to have rehearsed this eventuality prior to the commission of the crime. For the more the one convinces the other that under the circumstances just described she would keep the faith, all the more reason would there be - there being no honour among thieves, remember - for each of them to rat.

Any situation which satisfies the formal features of this story is called a PD. And a great many philosophers, myself included, believe such situations are sufficiently prevalent in the human interpersonal condition they have behooved our rationality to adopt strategies for extricating ourselves from just such sub-optimal results. Or - to put the matter in post-Darwinian terms - among organisms whose survival and wellbeing depend most heavily upon such interactivity, there will be powerful, natural, selective pressures in favour of those who manage to develop solutions to just such interactive impasses. [7]

Precisely three such strategies, or solutions, have emerged. And each, we believe, are the key to understanding, respectively, a) universalizing rationality, b) our political arrangements and obligations, and c) morality.

The first - which arises from the suspicion that the PD is only an apparent dilemma - is to suppose that, while leaving all the formal features of the putative dilemma entirely intact, we can nonetheless think our way out of it. Call this, if you will, the 'deconstructionist' solution [8], a version of which, the 'symmetry argument', we will be looking at in considerable detail in due course. [9] The second and third, the 'externalist' and 'internalist' solutions respectively, both acknowledge there is a dilemma - and therefore something must actually be done to escape it - but differ as to what that something might be.

4. The Logic of Leviathan [10]:

The Hobbesian (or 'externalist') solution is to institute an enforcer, or 'Sovereign'. In the above case, this means using the proceeds from an earlier, more successful, heist to retain the services of some non-participating third party, that service being that in the event the current heist is thwarted and the above situation arises, she will break the legs of anyone who rats, be it one set of legs now - i.e. if one rats while the other keeps the faith - or both sets of legs later - i.e. if they both rat and even if she has to wait three years to do it! More generally, then, the investiture of the Sovereign is thought to so lower the expected utility of 'defection' that both 'players' alter their preference orderings towards 'cooperation' and the dilemma is thereby dissolved.

But in addition to a) the difficulty in being able to trust the Sovereign, b) the limits on just how much we might be willing to pay her, and c) the implausibility of such an external mechanism alone being adequate to account for the actual extent of social, political and legal compliance, d) the 'compliance problem', as it has been called, is simply pushed back onto the act of instituting the Sovereign in the first place and thence to an infinite regress. And this is thought to be so because - and this is especially clear in the case of nuclear disarmament - it seems there necessarily exists a moment in the process of compliance at which it ceases to be rational for at least one party to it to complete it. [11]

As we will see, each of these four problems impose corresponding constraints on any externalist solution. But, as we will also see, these same problems re-emerge as requiring constraints on any internalist solution as well. Accordingly, let us take a moment to familiarize ourselves with them.

5. The Role of the Sovereign:

At the outset of Book Two of the Republic, Glaucon offers the following gloss on what we might call the 'naive' contractarian account of "the nature and origin of justice". "They say," reports Glaucon:

> that to do wrong is naturally good, to be wronged is bad, but the suffering of injury so far exceeds in badness the good of inflicting it that when men have done wrong to each other and suffered it, and have had a taste of both, those who are unable to avoid the latter and practise the former decide that it is profitable to come to an agreement with each other neither to inflict injury nor to suffer it. As a result they begin to make laws and covenants, and the law's command they call lawful and just. This, they say, is the origin and essence of justice; it stands between the best and the worst, the best being to do wrong without paying the penalty and the worst to be wronged without the power of revenge. The just then is the mean between two extremes; it is welcomed and honoured because of men's lack of the power to do wrong. The man who has that power, the real man, would not make a compact with anyone not to inflict injury or suffer it. For him that would be madness. This then, Socrates, is, according to their argument, the nature and origin of justice. [12]

Why is this account 'naive'? Because, Hobbes realized,

> decid[ing] that it is profitable to come to an agreement with each other neither to inflict injury nor to suffer it ... [and], as a result, begin[ning] to make laws and covenants,

takes us not one step closer to "neither inflict[ing' injury [or] suffer[ing] it". Hobbes too acknowledges

> [t]hat every man, ought to endeavour Peace, as farre as he has hope of obtaining it; and when he cannot obtain it, that he may seek, and use, all helps, and advantages of Warre, [and t]hat a man be willing, when others are so too, as farre-forth, as for Peace, and defence of himselfe he shall think it necessary, to lay down this right to all things; and be contented with so much liberty against other

> men, as he would allow other men against himselfe. [13]

But, unlike Glaucon, he saw all too clearly that this "com[ing] to an agreement", this "mak[ing] laws and covenants", and this "lay[ing] down [of] the right to all things", requires some mechanism of enforcement, a mechanism in the absence of which these "agreement[s], laws and covenants" would be about as effective as the aforementioned "rehears[ing] this eventuality prior to the commission of the crime. For," recall,

> the more the one convinces the other that under the circumstances just described she would keep the faith, all the more reason would there be - there being no honour among thieves, remember - for each to rat.

Clearly what is required, Hobbes saw, is "a common Power to feare". Indeed, he insisted,

> the Validity of Covenants begins not but with the Constitution of a Civill Power, sufficient to compell men to keep them. [14]

6. The Corruption Problem:

But whoever this "common Power" may be, if she be someone who is now "able to avoid the suffering of injury and practise the inflicting [of] it" by virtue of her 'clients' having "[laid] down their right to all things" - including, presumably, the means of resisting her enforcement of these "agreements, laws and covenants" - then one might be worse off than she would have been had she remained "in that condition of every man, against every man, which is called Warre." [15]

Nor is the problem of corruption - or what Samuel P. Huntington calls the problem of 'objective civilian control' [16] - confined to contractarians. For example, having taken pains to show that the account of "the nature and origin of justice" reported by Glaucon is wrong-headed, and having proffered a solution of his own - the realization of which requires guardians whose physical qualities include

> quick[ness] to see things and swift[ness] in the pursuit of what he has seen, and also [strength] if it is necessary to catch up with the enemy and fight to the end,

and whose mental qualities include, "bravery [and a] high-spirit ... fearless and unconquerable" - Socrates himself asks

> [h]ow, being of such a nature, they will not be savage in their behaviour to each other and to the rest of the citizens? [C]ertainly they must be gentle to their own people, but hard for the enemy to deal with. Else they will not wait for others to destroy the city but destroy it themselves first ... What shall we do then? Where shall we find a character which is both gentle and high-spirited at the same time? For a gentle nature is the opposite of a spirited one ... And surely if either of these qualities is missing, he cannot be a good guardian. The combination seems impossible, so it follows that a good guardian cannot exist. [17]

Of course ultimately Socrates solves the problem with recourse to what more modern sociobiologists call 'imprinting'. [18] And for Hobbes, of course, corruption is a 'problem' in name only. For since all that the investiture of the Sovereign need promise is a state of affairs preferable to "the condition of Warre" her investiture is designed to end, it becomes virtually impossible for her to fail to make good on that promise.

But political theorists after Hobbes - most notably Locke, Hume, the American Federalists, and Mill - thought we might expect more from our governments than just this marginal advantage over anarchy. That is, we can have the "Peace" we seek at considerably less cost in the currency of our vulnerability to the arbitrariness of the Sovereign than Hobbes thought. And, I suspect, contemporary bargaining theory would probably support them over Hobbes on this score.

But there is a second, and odder, point to be made here as well. And that is that, as much as we abhore the rampant corruption of the law, we are perhaps even less enamoured of bureaucrats so committed to the letter of it they are incapable of seeing beyond it to the purposes it purports to serve. In other words, the corruptibility of an enforcement mechanism can be a vice, but sometimes it can also be a virtue!

Just exactly how this observation manifests itself once we export it from the logic of external mechanisms to that of internal sanctioning remains to be seen. But even at this stage we can anticipate it will suggest a degree of tolerance and plasticity to our moral dispositions. Just how tolerant and how much plasticity, of course, we will have to discover.

7. The Expense Problem:

In a nutshell the expense problem is just this. For any given player the cost of enforcement must be less than - or in any event no greater than - the difference between the utility she could expect from mutual defection and the utility she can now expect from mutual cooperation.

Fair enough. But how often is this constraint satisfied? Often enough to warrant the investiture of the Sovereign, no doubt. But often enough too it is as Butch Cassidy said to the Sundance Kid: "If they'd pay me to stop robbin' 'em what they're payin' to stop me robbin' 'em, I'd stop robbin' 'em!" [19] Indeed, what motivates the anarchist - including the anarchist within each of us at one time or another - is that the cost of government regulation and enforcement can, and all too often does, outstrip the gain that such enforced cooperation purports to achieve. So just as post-Hobbesian political theorists saw we could have peace without absolute monarchy, they saw too we could have peace at considerably less cost to ourselves than Hobbes thought in the dual currencies of material wellbeing and forfeiture of liberty.

That said, Hobbes' core insight continues to inform even contemporary social and political debate. This is especially clear in the debate over the limits of freedom of expression. In a paper entitled "Why Dialogue?", for example, Bruce Ackerman urges us to place conversational restraints on interlocutors to avoid the discussion of contentious issues concerning the good. [20] The liberal emphasis - critics call it an obsession - with freedom of expression notwithstanding, Ackerman may well be right. If - as may well be the case in the debates over abortion and Shoah-denial, for example - such 'discussions' are likely to engender such hatred that one would rather risk being killed herself than suffer her interlocutor to live, 'defection', in game-theoretic terms, becomes once again the 'dominant' strategy. So provided one would rather be restrained from reaching such a state in the first place - on condition, of course, the other is similarly restrained - censorship can be countenanced, even under the liberal model. So, more generally, whereas liberalism can allow cultural self-criticism virtually without limit, what it can legitimately censor is such criticism as is likely to produce such divisions that 'defection-dominance' re-emerges.

Mutatis mutandis in the logic of internal sanctioning such limits on 'expenditures for peace' will manifest themselves instead in principles of computational economy. But, once again, more of this in due course.

8. The Compliance Problem:

There is considerable dispute among Hobbes scholars as to what he took his solution to the compliance problem to be; and then, within each reading thereof, further dispute yet about whether or not his solution succeeds. [21] My own reading is, I think, as exegetically supported as many, and certainly more charitable than most. But, as we will see, nothing very much apropos our project hangs in the balance, provided only either a) Hobbes' account fails, or b) it succeeds but not where it must were it to de-motivate our project. To wit:

Hobbes drew a distinction between what he called on the one hand 'Dominion' or 'Commonwealth' or 'Sovereignty' by 'Institution' and, on the other, dominion, commonwealth, or sovereignty by 'Acquisition', the latter differing from the former, said he,

> onely in this, That men who choose their Soveraign, do it for fear of one another, and not of him whom they Institute: But in this case, they subject themselves, to him they are afraid of. [22]

More specifically,

> [a] Common-wealth by Acquisition, is that, where the Soveraign Power is acquired by Force; And it is acquired by force, when men singly, or many together by plurality of voyces, for fear of death, or bonds, do authorize all the actions of that Man, or Assembly, that hath their lives and liberty in his Power. [23]

Now there are two views of the role of Acquisition in Hobbes' account. The 'received' view is that the essence of his story is Sovereignty by Institution. But Hobbes realized that insofar as neither the particular Sovereign currently in power nor any of his ancestors were as a matter of historical fact so brought to power, the would-be rebel might consider himself bound by bonds of fear only rather than by bonds of both fear and obligation; the crucial difference being that fear alone does not preclude rebellion whereas, or so Hobbes hoped, fear coupled with a sense of obligation does. Hobbes' answer to the would-be rebel, then, is that the Sovereign by Acquisition inherits the rights of the Sovereign by Institution, and does so more or less along the justificatory lines already suggested by our story of the two siblings and the pieces of cake.

Well, so far so good. But the key question is whether Hobbes thought the compliance problem for Sovereignty by Institution could *itself* be overcome. If so the received reading of Acquisition seems right. But if not Acquisition becomes the *only* means by which Sovereignty can come about. This is not to suggest that if, as it turns out, all Sovereignty is by Acquisition, then Hobbes has failed to ground obedience to the Sovereign. As we have already seen, a justificatory line can be drawn nonetheless. Nor is it to suggest that if Hobbes *wrongly* thought the compliance problem for Sovereignty by Institution could be overcome then his account of obligation would fail. It would succeed via that justificatory line drawn from Sovereignty by Acquisition. What it does suggest, however, is we can read Hobbes as himself acknowledging the compliance problem for Sovereignty by Institution cannot be overcome, while at the same time acknowledging his argument for obedience to the Sovereign could go through nonetheless.

This, then, is precisely how I prefer to read his concession that

> [i]t may peradventure be thought, there was never such a time, nor condition of warre as this; and I believe it was never generally so, over all the world. [24]

This is not to say we should ignore his observation that

> but there are many places, where they live so now. For the savage people in many places of America, except the government of small Families, the concord whereof dependeth on naturall lust, have no government at all; and live at this day in that brutish manner, as I said before. [25]

But what this says, and _all_ it says, is there are people in the world who have yet to achieve the benefits of Sovereignty. What it does _not_ say is they have failed to solve the compliance problem for Sovereignty by Institution, suggesting thereby there are people who have. Nor does it suggest these people someday _will_ solve the compliance problem for Sovereignty by Institution, rather than have to wait until it is solved for them by default, i.e. by Acquisition. So it remains entirely possible that what Hobbes has in mind here is that the compliance problem can _only_ be solved by Acquisition, and that these "savages" have simply failed, so far at least, to have the problem so solved for them.

What is key, however, is what follows. "Howsoever," Hobbes continues,

> it may be perceived what manner of life there would be, where there were no common Power to feare; by the manner of life, which men that have formerly lived under a peacefull government, use to degenerate into, in a civill Warre. [26]

For what Hobbes means here is not that this "perception" would miraculously enable people who are without a "common Power to feare" to put one in place, but rather and only that it should enable people who do have a "common Power to feare" to cherish that Power and to take precautions against its removal or degeneration.

Suppose instead Hobbes thought the compliance problem for Sovereignty by Institution could be overcome, and it could be overcome without there being already in place a meta-Sovereign, and so on *ad infinitum*. Then he must have thought - contra our claim (above) that the externalist solution to the PD is one distinct from what we have called the deconstructionist solution - that we can *think* our way out of, if not the PD proper, then at least one of the features constitutive of it. Now as we will see in due course, a precondition of complying in the absence of a meta-Sovereign with a commitment to cooperate in the investiture of a Sovereign is that the player in question has already replaced her SM-rationality with a disposition, at least when confronted with a PD, to cooperate with a co-player who she has grounds to believe will cooperate as well. So, if Hobbes thought the compliance problem for Sovereignty by Institution was solvable, he must have thought even "in a condition of meer nature" such a 'replacement' was possible. That is, not only must he have thought that "hav[ing] had a taste of both" "Warre and Peace" - or at least the wherewithall to imagine the taste of both - it would have occurred to one that it was SM-rational when confronted with a PD to adopt (what we will eventually see is) a 'symmetrist' rationality, but he must also have thought this aforementioned 'occurring' would be a sufficient condition for this aforementioned 'adopting'. But, as we have already seen, in the absence of virtual apodeicticity that others will adopt symmetrist reasoning as well, it would not be SM-rational to adopt symmetrist reasoning. And *a fortiori* it would be irrational to replace one's SM-rationality with self-constraint in the presence of such apodeicticity. So unless Hobbes thought one could think her way out of such 'dominance reasoning' - which would have made him a deconstructionist rather than an externalist - he could not have thought the compliance problem for Sovereignty by Institution could be overcome.

So, if we want to preserve Hobbes as an externalist rather than a deconstructionist, we should hold, if we can hold, that

1) Hobbes did not think - as Jean Hampton, for one, seems to think he did [27] - that non-compliance arises from the frailties of human reason and/or self-control, but rather - as I will be arguing -

that non-compliance is an ineliminable mandate of reason itself; that therefore

2) Hobbes did not think the compliance problem for Sovereignty by Institution could be overcome; that

3) he therefore thought as a matter of necessary historical fact all Sovereignty is by Acquisition rather than Institution; but that

4) obligation to the Sovereign arises out of our realization that, had the compliance problem as a matter of historical fact not been solved by Acquisition, we would all sorely wish it had been.

The problem remains, however, that one's sorely wishing the compliance problem had been solved by Acquisition is not the same - and Hobbes saw this too quite clearly - as her being capable of solving it by Institution. So, we might be tempted to ask, what kind of creature is it that at one and the same time has a) the intelligence to realize it is b) so lacking in self-control c) it needs an absolute authority over it notwithstanding d) it would resist that authority with every fibre of its being had it the power to do so notwithstanding e) it is grateful that it does not as a matter of fact have that power?!

The alternative view - and this seems to be Hampton's [28] - is to take this paradoxical psychology as highly problematic and therefore to suppose instead that

5) Hobbes did not believe that non-compliance is an ineliminable mandate of reason but rather, on his view,
6) it arises only from the frailties of human reason and/or self-control [29]; but that
7) he nonetheless thought the compliance problem for Sovereignty by Institution could be overcome; but
8) given the ineluctability of these frailties his account fails. [30]

But notwithstanding that mine is the more charitable reading, I grant that Hampton's is certainly among those which are, as I say, "as exegetically supported as any" and, perhaps, better supported than my own.

As we will see, the compliance problem re-emerges in the internalist solution to the PD in the guise of 'recidivism'. But more of that anon.

9. The Empirical Inadequacy Problem:

Now whether with recourse to Acquisition and Institution, or with recourse to Acquisition alone, with respect to the positive law, the police, the courts, and the government in general, I think Hobbes has the story more or less right. But clearly there is a further story that must be told if we are to account for those myriad instances in which there is no "common Power" - nor for that matter even dire disapprobation - "to feare", and yet we comply with our undertakings (and even our understandings) nonetheless. That is, there are rings of Gyges aplenty in our lives [31]; and yet few of us take advantage of such 'invisibilities'. Or if we do we do only seldom.

Now it might be thought the internal sanctioning we are about to hypostasize is meant to supplement the external ones we have already discovered and that the two, taken together, will be empirically adequate. And this, of course, is precisely our intent. But sometimes - and we will see examples of this from time to time - the cost of empirical adequacy is explanatory surfeit. That is, sometimes the explanans we provide is stronger, i.e. more immodest [32], than the target explanandum may require. Should we, on that grounds alone, reject the explanans and look for a more modest alternative? Clearly we should. But sometimes no such alternative is available. In which case we are saddled - sometimes embarrassingly so - with altogether more than we need. For example:

It is one thing to ask why, when you and I are eating together the local pizza parlour offers two for the price of one; and yet, when I am eating alone it refuses even to discount the one. No doubt it has some answer. But it is quite another to ask me why I order two, notwithstanding I am eating alone, I cannot eat two, and I have no intention of taking the second one home with me. Surely the question could as readily be, Why not? as, Why? And if the more natural question is, Why? rather than, Why not?, this can be so only because there is some suppressed background assumption, e.g. we are all committed not to waste resources regardless of whose resources they may be. But in the absence of such a background assumption would explanatory modesty settle the issue of which is the more appropriate question?

The moral of this story is that two dispositions - e.g. to order as much as one can eat and to order more - can be equally warranted, notwithstanding one may be more than what is warranted. But that warrant would be an adequate account nonetheless even for this excessive disposition provided only that excess involved no additional cost that might counsel against it.

This principle may prove important. So let us give it a name. To distinguish it from the principle of _unique_ explanatory warrant - which we reject - let us call it empirical adequacy's Principle of _Sufficient_ Reason. We will have occasion to appeal to it in due course.

3. THE LOGIC OF MORALS BY AGREEMENT [1]

1. Constrained Maximization:

As already indicated, it was to overcome, or at least ameliorate, all four of these difficulties - i.e. the corruption problem, the expense problem, the compliance problem, and the empirical inadequacy of external sanctions alone - that David Gauthier has proposed instead an *in*ternalist solution to the PD (and cognate 'games'), according to which the agent acts on the strategic rationality of inculcating in herself a disposition, if not transparent then at least translucent to others, to cooperate rather than defect with those, but only those, she has good grounds to believe will cooperate with her. [2] Or, in Gauthier's own words:

> Her disposition to cooperate is conditional on her *expectation* that she will benefit in comparison with the utility she could expect were no one to cooperate. Thus she must *estimate* the likelihood that others involved in the prospective practice or interaction will act cooperatively, and calculate, not the utility she would expect were all to cooperate, but the utility she would expect if she cooperates, given her *estimate* of the degree to which others will cooperate. Only if this exceeds what she would expect from univeral noncooperation, does her conditional disposition to constraint actually manifest itself in a decision to base her actions on the cooperative joint strategy. [3]

But, of course, of these 'good grounds' among the best would be that the co-player has translucently inculcated in _her_self the same disposition. That is, she

> makes _reasonably certain_ that she is among like-disposed persons before she actually constrains her direct pursuit of maximum utility. [4]

And, of course, of these 'good grounds' among the _worst_ would be that her co-player has translucently _failed_ to transcend her SM disposition. Thus,

> faced with persons whom she believes to be straightforward maximizers, [she] does not play into their hands by basing her actions on the joint strategy she would like everyone to accept, but rather, to avoid being exploited, she behaves as a straightforward maximizer, acting on the individual strategy that maximizes her utility given the strategies she expects the others to employ. [5]

Gauthier calls this disposition 'constrained maximization' or CM. [6] Since by adopting this CM disposition it is unlikely the player can do worse than she would otherwise - and yet it is certain she can do considerably better were she to happen upon a fellow CM-er - it should not surprise us that, or at least if, selective pressure has rendered most of us just so CM-disposed. [7]

2. _The Incoherence and Undermotivation Problems_:

I will have more to say in due course about this 'inculcating' and this 'translucency'. For the moment, however, I want to depart from my characteristic support of Gauthier to point out that the CM disposition, at least as Gauthier has so far characterized it, is both undermotivated and incoherent. It is undermotivated because these 'good grounds to believe' would surely include that one's co-player has translucently adopted a disposition to cooperate _unconditionally_. And yet it would clearly be irrational to cooperate with _her_. And it is incoherent because if the CMer must

> make reasonably certain that she is among _like_ disposed persons before she actually constrains her direct pursuit of maximum utility,

then constrained maximization is trapped in an infinite regress of mutual conditionalities.

3. _Danielson's Artificial Morality_:

Both of these points are made by Peter Danielson in his recent _Artificial Morality_ (_AM_). [8]

Gauthier and Danielson are of a mind that (what Danielson calls) the "fundamental" justification for morality - that is, one "that does not appeal to any of the concepts of [morality itself]" [9] - is that, conventional wisdom notwithstanding, nice guys - though, adds Danielson, not necessarily the nicest guys - finish first! But where _Morals by Agreement_ (_MBA_) "borrows techniques from game theory to reduce social problems to their abstract essence" [10], AM supplements these techniques with

> tools borrowed from artificial intelligence. [It] construct[s] artificial agents that interact in a small, toy world called H*Land, [a world] roughly characterized by three features: First, it promises to return benefits to those who cooperate to exploit its natural resources. But since it is initially barren of moral[ity], the returns to cooperation are only potential; agents may not cooperate and cooperators may be exploited. Second, a logically precise description of each player suffices to generate its described behaviour. And third, it is a virtual place; only software entities are permitted. [11]

'Players' are reduced to mere "decision functions that return either 'cooperate' or 'defect' when asked to make a move in a game." [12] And these 'decision functions', in turn, are modelled "in the computer language of Prolog." [13]

Furthermore, whereas Gauthier takes considerable pains to "relate [his] idealizing assumptions to the real world" [14], Danielson, by contrast, prefers "to understand the 'logic' of cooperation and constraint in clear cases before wading into messier practical approximations." [15] And so where MBA claims its "premises are true of people" [16] - and therefore what it offers us is sound advice about what moral dispositions to adopt - AM, by contrast, purports to "take its source of constraint only from the limits of what is procedurally possible" [17], and so confines itself to mere validity. [18] Still, claims Danielson, "programming artificial players to score well in tournaments of various players playing abstract non-iterated mixed motive games is a good way" [19] to discover - if not which moral dispositions are, as it happens, 'fundamentally' justified in the real world, then at least - how one goes about 'fundamentally' justifying a moral disposition, wherever one might find herself. And so, claims he, "players successful in [this] world may teach us something [at least] about how to deal with our [own]." [20]

Accordingly, save for states and corporations, Danielson does not claim that one can simply map the lessons learned in H*Land onto the somewhat 'messier' human condition. [21] For, notes he, unlike H*Land entities,

> we humans are not cognitively transparent, we lack the discriminating means of commitment that rational morality requires, and we cannot readily adapt our commitments, as our emotional mechanisms for fixing dispositions tend to have high enertia and momentum. [22]

Indeed, I would add, there are, as any cognitive scientist will attest, good reasons why we are as cognitively opaque, indiscriminate and recalcitrant as we are. For one thing, in the human world windows of opportunity for profitable cooperation and/or the avoidance of life-threatening exploitation typically expire in a matter of seconds. In other words we do not have the time to indulge in the kind of algorithms that virtual entities can perform. And for another, the power of our 'wetware' is a function of an equilibrium long since negotiated with the burdensomness of its supporting boneware and executive musculature. In other words we do not have the brains to indulge in the kind of algorithms that virtual entities can perform. These reasons, then, are as endemic to the logic of human cooperation and constraint as any other feature of the 'original position' from which, on this contractarian model, our moral dispositions are 'fundamentally' justified. Thus to suppose one can fully capture the logic of human cooperation and constraint without reference to these further reasons is tantamount to supposing one can capture the rationale behind the programming strategy for a particular chess-playing machine without reference to the response time beyond which the machine will be deemed to have forfeited the game and the computational and storage capacity of the hardware on which the program is to be instantiated.

In short, human - as distinct from virtual - rationality demands compromises even with itself.

That said, Danielson cannot be faulted for failing to consider some of these messier complications. Had he essayed a treatment of all the variables to be factored in a comprehensive logic of human cooperation and constraint, he would have had to write a library rather than a book. And, to be fair, he does factor in response-time and hardware-burden, albeit, as we will see, under the veiled rubric of 'information costs'. [23] What remains to be seen, however, is whether this takes response-time and hardware-burden seriously enough.

4. Campbell's and Danielson's Solutions to the Incoherence Problem:

By 'CM' Gauthier intends what Danielson calls 'conditional cooperation' (CC); that is, "cooperate with and only with those who one expects to cooperate." [24] And, as we have already seen, thus formulated constrained maximization is incoherent because it traps itself in an infinite regress of mutual conditionalities, and it is undermotivated because it sub-optimally forgoes opportunities to exploit 'unconditional cooperation' (UC). [25] Let us deal with the "coherence" or "looping" or "non-circular assurance" problem first.

The 'received' solution to Gauthier's coherence problem is that proffered by Richmond Campbell. Campbell proposes

> we modify the form of Gauthier's definition of CM so that it has the following structure: When player S (self) and O (other) are in a PD, S has the CM-disposition iff: 1) S has the property R and 2) S will cooperate with O iff S believes that O has R.

For example, suggests Campbell,

> let R be the property of having a red nose. Certainly no circularity would exist given _that_ specification of R. Each of any two CM's would have a red nose and, provided that they could see the colour of the other's nose, they would cooperate in the PD. In general,

then, concludes Campbell, "there is no circularity as long as the specification of R does not make any reference to how the other party will act on _this_ occasion." [26]

The 'received' _objection_ to Campbell's solution, in turn, is that proffered by Holly Smith. One _cannot_ "avoid both circularity on the one side and exploitation on the other," she opines. [27] Why? Because, though

> [i]t is true that SMs won't have the property Campbell describes, so that someone who adopts CM will not be exploited by SMs ... there will be other defecting players who _would_ reciprocate cooperation in sequential [PDs], but who would _not_ cooperate in simultaneous [ones]; an agent adopting Campbell's version of CM would be vulnerable to exploitation by these players. [28]

What kind of players might these be? Since Campbell's player might aptly be called 'Rudolf', Danielson (staying in season) invites us to consider a character as aptly named 'Grinch'. [29] Grinch employs CC whenever he moves first in a sequential PD, but in the simultaneous one he reverts to 'self-same co-operation' (SC); that is, "[he] co-operate[s] with and only with players similar to [him]self." [30] O is 'similar' to S, says Danielson, just in case, having exercised their "right [to] reveal" their decision procedures to each other [31], "[S] can swap [her own] name for [O's] in [O's own decision] procedure and get [S's] own [decision] procedure." [32] Thus, since Rudolf is dissimilar to Grinch notwithstanding their like-coloured noses, Rudolf will cooperate while Grinch defects.

To block this possibility Danielson proposes the following. First we insist on access to the other player's decision procedure. Next we "link indirectly to the other player's action by mentioning her decision procedure" in [our] own decision procedure. And finally we give her decision procedure access to our own. [33] Quad erat demonstrandum. The coherence problem is solved!

5. Danielson's and Campbell's Solutions to the Incoherence Problem Revisited:

When a problem that has plagued philosophers for two and half millenia is solved in the span of a sentence, chances are something has gone unnoticed. And indeed something has.

What has gone unnoticed is that this pseudo-CC (PCC), having just unilaterally defected on CC-proper, can herself be unilaterally defected on in turn by pseudo-SC (PSC). PSC pretends to give her co-player access to PSC's own decision procedure - by revealing everything up to and including her 'execute' command. [34] But in fact she holds back on a secret protocol, mole or virus, which allows her to defect the moment she is assured - or at least she thinks she is assured - her co-player will cooperate. In short, what Danielson does not explain is how S assures herself that what O has revealed to her is not just "the truth and nothing but the truth" but also "the whole truth" about O's decision procedure.

Suppose even that S queries herself as to where in her own decision procedure she'd be likely to hide such a protocol herself and then insists that O open the corresponding location in O's decision procedure to S's inspection. Still, if there is an infinite number of addresses in which one could hide such a virus, it follows by Cantorian diagonalization there forever remains an address that has yet to be queried and inspected. Nor can S and O overcome this infinite regress - and hence the problem of non-circular assurance - by transparently rendering themselves computationally finite. For each would then need assurance the other was not holding out on this 'rendering', or complying with the rendering but holding computational space in unreported reserve.

More generally, then, to suppose S and O can solve the assurance problem by rendering themselves procedurally transparent to each other is to suppose they have already solved the assurance problem for rendering themselves procedurally transparent to each other. And the only way they can do that, I submit, is for Danielson to make it a further feature of H*Land tournaments that:

> each of its contestants have a priori knowledge of the address capacities available to each of her co-contestants.

But if, in order to salvage his solution to the coherence problem, Danielson is prepared to add this feature, then he and Smith are no longer in a position to so readily dismiss *Campbell*'s solution. Why? Because by "property R" what Campbell had in mind - or, if not, then what he *should* have had in mind - is:

> any external property the having and/or maintaining of which is known - either *a priori* or by experience - to go a reasonably reliable distance towards exhausting the computational space remaining in which one might hide any such 'feint and deek' protocol.

A case in point of such a property - and I suspect something more like this than red-nosedness is what Campbell had in mind - is the property of "seeming to wear one's cooperative dispositions on one's sleeve." One certainly *can* manifest such a property disingenuously, i.e. while simultaneously *not* wearing one's cooperative dispositions on one's sleeve. But - as those of us who have attempted to do so can attest - maintaining the ruse is literally exhausting, it is more often than not unsuccessful and, in the final analysis in any event, the costs incurred *by* such 'passing', even if successful, almost invariably outweigh the benefits that ac-crue from it. That is, one loses more in the time and effort it takes to *feign* trustworthiness than she can hope to gain by the feint. And, moreover, that this is so is known to her co-players.

Of course we have long since known this to be case in the human world. What we have just discovered is that it is equally the case in H*Land. What is more we have discovered why. What we have discovered is that the one and only solution to the coherence problem is to leave any and all 'feint and deek' protocols no place to hide. In the human world this is done by time limitations and hardware burdens. In H*Land these constraints are unavailable. And so there it must be done instead by loading the game with a priori knowledge of the extension of each player's internal playing field. But, it is to be noted, once Campbell's solution is thus clarified and Danielson's thus repaired, his and Campbell's solutions turn out to be one and the same.

6. Gauthier's Solution to the Compliance Problem:

Danielson attributes the second of Gauthier's errors to his failure to extend his own insights into the logic of dispositional pluralism beyond the mere dualism of SM and CC. He honours Gauthier as among the first to realize that the disposition with which it is rational to enter a population - and/or to adopt once in it - is a function of the make-up of the population in question. But, he notes, what Gauthier fails to take seriously is that the set of available dispositions is by no means exhausted by SM and CC. Neither in turn, therefore, is the set of possible population make-ups a function of the interplay of only SM and CC. Rather it is a function of the interplay of SM and all programmable variants of CM, including, as already noted, UC, but including as well reciprocal cooperation (RC) - i.e. "cooperate when and only when cooperation is sufficient and necessary for the other's cooperation" [35] - not to mention the innumerable variations in between.

For simplicity let us confine ourselves to CC, RC, SM and UC. And let us verify the observations that follow by adopting the simple reverse-ordinal scoring system, that being, recall: C/D=1/4, C/C=3/3, D/D=2/2, and D/C=4/1. Thus in the general case - of which any of the simpler ones to follow can be treated as merely 'special' - where c, r, s and u are the numbers of CC, RC, SM and UC respectively, one scores as follows:

$$CC = 3(c+r+u)+2s-3$$

$$RC = 4u+3(c+r)+2s-3$$

$$SM = 4u+2(c+r+s)-2$$

$$UC = 3(c+u)+r+s-3$$

We need now only add that:

invasion: A new disposition is said to 'invade' an existing population just in case it does as well in the course of the attack as the member of the existing population that performs best.

dislodge: A disposition is said to 'dislodge' another just in case the first fares *better* than the second.

superiority: A disposition is said to be 'more substantively rational than' another - or, less cumbersomely, 'rationally superior to' it - just in case "the first could invade a population consisting of the second and [yet] the second could not invade [a population consisting of] the first." [36]

And, finally,

equilibrium: A population is said to be in 'equilibrium' just in case "a new entrant [of an extant type] receives the same payoff regardless of its type." [37]

With this apparatus in place let us begin with the simplest case, i.e. Gauthier's. In a tournament consisting of only two contestants, CC and SM, neither can resist invasion from the other, but neither can either dislodge the other. That is, in the absence of another CC CC fares no worse than SM, but neither does she fare any better. Nor does this change even if a singleton CC is facing a veritable phalanx of SM. But in a contestant-pool of more than one CC CC invades and dislodges SM, and does so irrespective of the number of SM. Moreover, once two CC have taken the field, SM can neither dislodge nor invade. In other words, since it can never hurt it is rational to adopt CC (even if only) on the off chance of happening upon another CC.

This in a nutshell is Gauthier's argument for CC. But now let us move on to some of Danielson's innovations.

7. The Rationale Behind the Re-Definition of Superiority:

We have already seen that we need not confine our jousts to single combat. We can also conduct tournaments-proper, i.e. in which there may be more than one token of the attacking type attacking a field of more than one token of the defending type defending it. But, notes Danielson, even "this simple picture is complicated [were we to consider] the presence of third parties. [For, as we will see, such] intermediaries can change the outcome." [38]

For example, in a tournament consisting of any number of CC and UC, neither can resist invasion from the other but neither can either dislodge the other. That is, in the absence of even a single SM UC fares no worse than CC; but in the absence of even a single RC neither does CC fare any better than UC. Moreover, though neither CC nor RC can dislodge the other, neither can either resist invasion from the other, since in the absence of UC neither fares better nor worse than the other. Nor in the absence of UC can SM ever invade (save, trivially, for a homogeneous population of itself). So the only heterogeneous populations that are internally zero-tolerant of defection, and therefore internally stable, are CC/RC and CC/UC.

But now let us see what happens when these communities are attacked. Though neither CC/RC nor CC/UC can be invaded by SM, nevertheless in the case of the latter, i.e. CC/UC being attacked by SM, UC will be dislodged in the attack along with the attacker. UC will dislodge CC in its own doomed attack on CC/RC. And in attacking CC/UC RC will replace UC and CC. So SM (albeit suicidally) 'ethnically cleanses' CC/UC of UC for CC; UC (albeit suicidally) rids CC/RC of CC for RC; and RC runs genocidal amok on both CC and UC.

But, notes Danielson, "we don't want to test a strategy in the presence of irrelevant intermediaries." (What he means by 'irrelevant' will be made clear momentarily.) So, he continues, "we need to develop a test that determines the drift to intermediaries as well as deciding rational superiority in terms of invasion." [39] To which end he proposes "we start with a population" consisting homogeneously of the disposition "to be tested". Next we add to it 'non-arbitrary additions', by which he means "[any] player [that] does as well as any member of [this] existing population." The population that results is then deemed to be a 'rational extension' of the population to which the non-arbitrary addition has just been grafted. We then apply this function recursively. And, finally, we revise the definition of 'rational superiority' (given above) to read:

superiority*: A disposition is said to be 'more substantively rational than' another just in case the first can invade some rational extension of the second, and

> yet the second cannot invade any rational extension of the first. [40]

What is the force of this 'some' and this 'any'? And what is the relation between this revised definition of superiority and the notion of 'irrelevance' that inspires it? To make a long story short, Danielson is confessing here that at this point in his thinking he faced "a choice point". He could have strengthened the conditions for superiority. That is, he could have acceded to Holly Smith's "demand [that] instrumental rationality [manifest] what we might call a parametric robustness, an ability to do well against [a] wider variety of strategies, rational or not." [41] Instead, he reports, he elected a "standard of instrumental success [that] stresses the strategic model of rational choice, i.e. what works best among the [at least] (weakly) rational." [42] Had he "follow[ed] Smith's lead," he explains, he would have had to "introduce an unlimited variety of players and [so] quickly overwhelm [his] ability to manage complexity and advance our understanding of the issues." [43] So, in layperson's terms, the position Danielson is taking is just this. Interactive rationality can require us to be able to fend off players who in their own interests are legitimately trying to beat us at our own game. But it cannot require us to be able to fend off players who have not a hope of beating us at our own game, and have thrown themselves against us for the sole purpose of frustrating in our interactions with those who are legitimately trying to beat us at our own game.

And why is Danielson so keen to make this point? Because the only way he can argue for the superiority of RC over CC is by including UC in the substantive rationality tests. But, his detractor might object, if Danielson can

> include contrived king-makers like UC to bias the test in favour of RC and against CC, why [can she, the detractor, not] add contrived king-breakers [to reverse the bias]? For example, [why can she not] introduce an agent who refuses to cooperate with RC, thus making RC do worse and CC do better, [thereby] undermining [Danielson's] claim for the substantive superiority of the former? [44]

The detractor now has her answer. The difference between RC's king-maker, UC, and RC's king-breaker - shall we call her Shaft-RC (SRC)? - is that UC is a legitimate contender in her own right whereas SRC's sole raison d'etre is to make life miserable for RC.

8. Danielson's Solution to the Undermotivation Problem:

Of course what we have yet to consider - nor, here at least, shall we - are the effects of replacing ordinal scoring values with real ones. Nor need we complicate our story even further by factoring in prudential considerations. For Danielson's general case against Gauthier is already made. How so? Because, by Danielson's reckoning, from a homogeneous population of CC it will take precisely 32 recursions for the rational extension of this population to reach equilibrium. And the equilibrium it will reach will be 7 CC to every 2 UC to every 1 RC. [45] And, adds Danielson, "if scrutiny costs are higher, the equilibrium population may include SM as well." [46]

What this shows, of course, is that to the question, With what disposition would it be most rational to enter society? there is neither a categorical nor a universal answer. It all depends on the dispositional mix that makes up the particular society one is readying herself to enter. But what it also shows is that to the question, With what disposition would it be most rational to enter a particular society? there need not even be a unique answer. In fact, in the case at hand, any of CC, UC and RC will perform equally well. And yet Gauthier has claimed that "equal rationality demands equal compliance". [47] So on this very important score Gauthier is clearly in error. [48]

That said, beyond this de-privileging of CC are there any further insights being offered here? Well for one thing we can now see that there is a sense in which CC, by virtue of its unnecessary kindness to UC, is being transitively complicitous in the rewarding of SM. So if - as is most often the case - there is a limited plasticity to the size of the population, we _all_ have a vested interest in keeping to a minimum the number of SM among us. True, CC and RC, unlike UC, need not worry about being exploited by SM. But the point is that neither CC nor RC nor UC nor SM will be able to reap the benefits of mutually _cooperating_ with SM. But, as we have just seen, we rid ourselves of SM only at the expense of first ridding ourselves of UC; we rid ourselves of UC only by first ridding ourselves of CC; and yet CC cannot be got rid of without UC first being got rid of. So though we cannot establish that RC is any more than _as_ substantively rational as CC, we have at least shown that RC is _prudentially_ superior to CC.

9. _The Implications of Prudential Superiority_:

More generally, then, what _AM_ is offering us is a set of tools that can make good on a promise of incalculable explanatory, predictive, and manipulative power. With them in hand we are in a position to understand not only why certain domain-of-action-specific dispositions enjoy the mix they do but, as well and more importantly, what kinds of attacks, be they rational or not, might alter that mix in a direction more to our liking. For example:

AM's account of rational superiority allows us to see why the game of Subway Mugging is played out as so often it is. Divide the passengers into four personality types: hawks, doves, cowards and heroes. Hawks prey on doves. Heroes defend them. And cowards cower in a corner, to be noticed by hawks only when they run out of doves to victimize, at which point the coward *must* show her feathers. Of course heroes pay a higher price for their heroism than do doves for their passivity. Doves lose their purses, heroes (sometimes) their lives. So it pays to be a dove, provided there are enough heroes. It pays to be a hawk, provided there are not too *many* heroes. But it almost never pays to be a hero. The fewer heroes, however, the more hawks. And the more hawks the fewer the doves with purses for the snatching. The problem with the New York subway system, then, is that there just are not enough heroes. And, thanks to Danielson, we now understand why.

But what *prudential* rationality allows us to see are two things. First - though perhaps we should call this _advocacy_ rationality - though it may not be rational to be a hero oneself, it is certainly rational to encourage others to be. This is why it is almost invariably greyer and wiser men who are most keen to send younger and less wary ones off to the deserts of Kuwait. And second, though it may not be rational to show one's feathers today, it may be _prudent_ for a coward to become a hero now, rather than wait until the number of hawks is so high that the number of doves is so low that the coward must become either an undefended dove or a lonely hero. Which is why maybe the young men who went off to the Gulf were not so unwary after all!

Indeed, the modelling of a _prudential_ rationality, grafted onto the modelling of rational superiority provided by Danielson, is precisely the direction I think game-theoretic research should next be going!

But I digress. For I come, recall, not just to praise Danielson but, if I can, to raise some doubts about his claim that _AM_ "takes its source of constraint only from the limits of what is procedurally possible." [49]

10. The Definition of Superiority Revisited and, Thereby, a Revisit Too to Danielson's Solution to the Undermotivation Problem:

Danielson bootstraps his definition of superiority, recall, with reference to rational extensions of homogeneous populations of dispositions to be tested. But, it is to be noted, there is an infinite number of procedurally possible dispositions. So, one might well ask, what constrains the set of dispositions that AM takes itself obliged to test? If the answer is, Nothing at all!, then the answer to the question, What can we *learn* from AM?, will likewise be Nothing at all!, since it follows once again by Cantorian diagonalization that there forever remains a disposition yet to be tested, and therefore any claims about the superiority of a *tested* disposition remains forever merely provisional and therefore inconclusive.

Nota bene the question I am asking. I am not asking, Why are we *only* testing CC, UC, RC and SM? I know perfectly well we cannot test *every* procedurally possible candidate. Rather I am asking, Why are we testing *these* rather than some others? Note too that Danielson's notion of non-arbitrariness does not address this question. It determines only what *intermediary* dispositions should be allowed in our tests on dispositions entitled to testing in their own right. But what I am asking is, How do we select the dispositions entitled to testing in their own right in the first place?

Now it _might_ be argued - though I do not think successfully - that there is a kind of _natural_ privileging procedure available here, beginning with SM.

Premise 1: Most encounters with the world are parametric. (Plausible, but only if we are talking about the real world.)

Premise 2: SM is the appropriate disposition for parametric encounters. (No, SM is just one of an infinite number of dispositions appropriate to parametric encounters. It is, however, the least time-consuming and the least computationally burdensome. For example, "SM, whistle three bars of the Moonlight Sonata, _then_ execute!" is as appropriate to parametric encounters as "SM-simpliciter!" It just takes longer and requires an extra calorie or two to run it!)

Premise 3: The appropriate meta-disposition by which to convert from one context of reasoning to another is to start with what already comes naturally. (Perhaps, but, once again, only if the organism is worried about response-time and hardware-burden.)

Therefore: SM is the most natural candidate for a first test. (Yes, but notice how much has already been imported from the real human condition!)

But even if we do privilege SM, why CC? Why UC? Why RC? Are the dispositions to be tested those and only those that have made it into populations which are rational extensions of SM? No, because, as we have just seen, there is an infinite number of such dispositions. So, as a first pass at least, let us say that the dispositions to be tested are constrained by the dual principles of speediness and address efficiency. What we are looking for, we might suppose, are the most elegant programs up to the tasks at hand.

Fair enough. Now all we need to ask is, What constrains the tasks at hand? Presumably to fare better than SM, then to fare better than whatever fares better than SM, then on to faring better than that, and so on. But since at any given juncture in this process there is an infinite number of candidates, each of which might prove successful, are we forever condemned to merely picking at random and hoping for the best? Or is there, in fact, something that is guiding our selections?

Obviously there is. And, just as obviously, what it is is our own human experience. Why did Gauthier think to test CC? Was it because he could see immediately it might be superior to SM? Well, perhaps. But more probably because he observed something (at least) akin to CC in the world around him. Why did he think to test UC? Was it because he could see that in a hypothetical world of CC UC could take a free ride on its own lower scrutiny costs? Again, perhaps. But more likely because he saw around him human agents whose cooperation was not conditional, and so he thought to ask himself how such people could possibly survive? And why did Danielson think to test RC? Was it because he noticed around him people identical to the CC's noticed by Gauthier save that they are prone to defect on those UC's likewise noticed by Gauthier? Or was it because he was casting about for a disposition that might do better than CC and merely invented RC to do the job? Here the answer is not so clear. What is clear, however, is that insofar as the set of rationally non-arbitrary additions to a population is defined in terms of the homogeneous disposition-type to which these additions are grafted, we cannot define the dispositions to be tested in the first place in terms of rationally non-arbitrary additions to itself. That would be circular, and hence incoherent. What we need, then, is a non-circular meta-test for test-worthiness. And yet this is precisely what AM fails to provide us.

If we confine ourselves, as tentatively suggested above, to dispositions found in the real world, we run two risks. First, we might be missing out on the explanatory power of positing dispositions that may have been rendered extinct by subsequent improvements. For example, were there no longer any residual psychopaths we would have no way of guessing we were once the selfish 'bastards' portrayed to us by Hobbes. But second and much worse, we would be precluded from testing possible (but as yet untried) revisions to the way we are. AM would be descriptive, but neither reconstructive nor visionary.

I do not think there is a solution to this problem. Certainly we are entitled to test extant dispositions. And, just as certainly, we are not entitled to waste valuable research time on dispositions that simply cannot be run on human wetware or, even if they could, would be too slow to be of any use to us. Beyond these two constraints, however, I think we are condemned to just taking shots in the dark. We are, I think, in much the same position as Darwin's deistic deity. Having gone fishing on the Seventh Day, She left it to random mutation to fix the handiwork She had botched on the Sixth. Obviously someone somehow thought of something to improve upon our Hobbesian psychology. But if we look to _Artificial Morality_ to artifice our next quantum leap, we ask of it something that exceeds its limits!

11. _Constrained Maximization Repaired_:

So, back to our core project.

Imagine, then, there is a pill on the market - let us call it a Grinch-Buster! - that has the following three effects. First, it produces a particular mark on one's fore-head, a mark that can only be produced by ingesting this particular pill. Second, it hardwires her to cooperate - simultaneously _and_ sequentially - with those, but only those, with just such a mark on their foreheads. And third, it hardwires her _not_ to succumb to the temptation of taking an antidote when transferring over from a sequential game to a simultaneous one. Could such an agent ever be exploited? Apparently not. Are there opportunities for exploitation such an agent has forgone? _A fortiori_ not. So now we _can_ say that, by taking such a pill,

> the player can do no worse than she would otherwise and yet considerably better were she to happen upon a fellow Grinch-Buster.

And it is this disposition, I submit, that Gauthier should have identified as constrained maximization.

That said, I do not deny that in the course of the real world dialectic of disposition-acquisition Gauthier's version of constrained maximization might ultimately emerge. That is, since, as we will see, in the real world we have difficulty enough identifying our co-players, the predatory gains to be made from being able to distinguish between CC and UC may not be worth the loss in computational economy that such a recognition-module would involve.

12. Taking the 'Pill':

Of course in the real world the role of the 'pill' is played by operant conditioning and/or natural selection; and it is not so much taken as rammed down our throats - on our mothers' knees, in the sandbox, in the pew, in our genes. In other words, as a matter of historical fact it may turn out we have become moral by Acquisition rather than Institution.

But - to return to the aforementioned 'key' question - can Gauthier solve the compliance problem for constrained maximization by Institution? And the answer is: yes he can. Recall that for the Hobbesean rational maximizer it is irrational for Column to replace her default, SM, dominance reasoning with constraint - and this replacement remains irrational for her with or without assurance that Row has done likewise - unless both a) Column's disposition to cooperate can be made conditional upon Row's being similarly disposed and b) Row's disposition to cooperate is conditional upon Column's being so disposed. But it is precisely these conditionalities that the Hobbesean agent in a state of nature cannot count on.

The (repaired) Gauthierian rational maximizer, by contrast, has that assurance of mutual conditionality. She has it by virtue of the transparency condition. Or, as Gauthier prefers, she has at least virtual assurance of it by virtue of the translucency condition. So the essence of Gauthier's internalist solution to the PD, it seems, lies not so much in the CM disposition being internal to the agent but rather in cooperation being a) conditional upon a transparent (or at least translucent) trigger and, upon the satisfaction of that triggering condition, being b) hardwired. So if both a) transparency (or at least translucency) and b) conditional hardwired cooperation could be achieved by some external mechanism, that solution would nonetheless count, for our purposes at least, as an internalist one.

13. Translucency:

But why settle for mere 'translucency' rather than insist on much stronger, and therefore much safer, 'transparency'? Because, Gauthier explains,

> to assume transparency may seem to rob our argument of much of its interest. We want to relate our idealizing assumptions to the real world. If constrained maximization defeats straightforward maximization only if all persons are transparent, then we shall have failed to show that under actual, or realistically possible, conditions, moral constraints are rational. We shall have refuted the Foole but at the price of robbing our refutation of all practical import. [50]

Still, one might wonder, why is transparency too much to ask "under actual, or realistically possible, conditions"? For surely if cooperation were conditional upon mutual recognition of a transparent, rather than merely translucent, feature of the other, cooperation would be far less reticent, hence much more automatic, hence much more computationally efficient.

The answer, it seems, lies in the countervailing survival and self-promotion value of maintaining (preferably) non-detectable stake-thresholds at which one will toggle back to SM. At just what height rationality can set those thresholds will depend, of course, on in just what circumstances the organism can reasonably expect to find itself situated and the computational economic trade-off constraints on fine-tuning such thresholds to such situations. But the core insight is just this. That people can and do throw themselves on grenades to save their fellows cannot be denied. (I know because I have seen it.) But whether or not we are entitled to judge such behaviour as supererogatory, and hence irrational, will depend on whether or not there was an alternative self-preservation-threshold available that nonetheless would have allowed him to, say, at least risk a sniper's bullet in order to rescue a wounded comrade from open ground. (I have seen that too.)

Moreover, making a soldier out of a civilian involves, among other things, adjusting just those thresholds. But this should not surprise us. Human beings, being social beings, are specialized beings. So here is one question. Is it rational for us to so adjust him? Here is another. Is it rational for him to allow himself to be so adjusted by us? But here is a third which, for our purposes at least, can remain open. Are these two separate questions or just one?

14. The Domain of Discourse:

So far, it is to be noted, we have confined ourselves, as does Gauthier by and large, to the PD. But constrained maximization is not confined to a solution to the PD; nor is morality - in the attenuated sense employed by our schema - confined to constrained maximization. So let us cast our net a tad wider, if only to familiarize ourselves further with the pattern of explanation our schemata purport to offer.

Second only to the PD - in terms of journal tome-age at least - has been the Deterrence Dilemma (or DD). Suppose I am the American President at the height of the Cold War, and, appearances notwithstanding, that I am a reasonable and decent enough fellow, such that whereas I prefer American lives over Soviet lives, I prefer Soviet lives over no lives at all. And suppose too these facts about my preferences are known to my Soviet counterpart. This being the case, I can reason that he can reason that, had he sufficient cause to do so, he could provoke without fear of retaliation. And so this being the case, I reason, I should bring myself, if I can, to prefer no lives at all over Soviet lives, and to so prefer transparently. Moreover I should do so - notwithstanding that in the event this strategy fails and the Soviets launch nonetheless, I will have hardwired myself to do the monstrously irrational - provided only the expected disutility of retaliation (i.e. the direness of its consequences multiplied by its probability) is less than the expected utility of non-provocation (i.e. the benefits of _its_ consequences multiplied by _its_ probability). Moreover - for what it may be worth - I will _not_ in fact be acting irrationally when I retaliate, since at that time, _ex hypothesi_, to Soviet lives I _will_ prefer no lives at all.

Such DD's - or what Gregory Kavka calls 'Special Deterrence Situations' (or SDS's) - are enabling of a number of philosophical insights, not the least being that they seem to defeat both the Wrongful Intentions Principle (or WIP), i.e. that

> [i]f it would be wrong to do something under certain conditions, then it is wrong to form the intention to do that thing should those conditions arise [51],

and the Right-Good Principle (or RGP), i.e. that

> [d]oing something is right if and only if a morally good person would do the same thing in the given situation. [52]

For the RGP, it would seem, is defeated by the observation that in certain situations, i.e. SDS's,

> it would be morally right for a rational and morally good agent to deliberately (attempt to) corrupt [her]self [53],

and, in these same situations,

> it would be morally wrong for a rational and partly corrupt agent to (attempt to) reform herself and eliminate her corruption. [54]

Most important among what is highlighted by SDS's for our purposes, however, is the logical genesis of revenge. For whereas revenge is, by definition, the expenditure of a utile for no other purpose than to inflict a disutile on another - and therefore always itself irrational - the transparent disposition to seek such vengeance, conditional of course upon provocation, can be rational just in case the utile it protects, multiplied by the probability it will protect that utile, is greater than the utile conditionally expended, multiplied by the probability of the condition triggering its expenditure being met.

The more general case for the rationality of irrationality, however, is made out in turn by the Extortion Dilemma (or ED). You, the extortionist, break into my house, seize my child, place a pistol to his head and demand, "The combination to the safe or the kid gets it!" Now note that your pistol, albeit necessary, is insufficient to establish your power over me. Presupposed by you too must be that I value my child more than my money and that I am sufficiently rational to behave accordingly. Both, let us suppose, are safe enough assumptions on your part. But, I might reason, might there be a way I can keep my child and my money?

Certainly there is. Suppose there is a pill on the market - let us call it an ED:P1 - and hence in the candy dish on the coffee table - I keep them there for just such eventualities! - that produces the following two effects. First, it produces a mark on my forehead - a mark that can only be produced by taking just such a pill - and second, it instantly, albeit temporarily, renders me utterly irrational. "Yes, I love my kid," I prattle maniacally, "so go ahead and blow his brains out!" I am assuming, of course - but let us suppose rightly - that you, in turn, would rather turn tail and run from a charge of attempted extortion than from one of gratuitous murder. Now then, would it be rational for you to follow through on your threat? Clearly not. But, you might reasonably reason, might there be a way for you to get the money after all?

Clearly there is. Having been thwarted by this kind of pill before, you have come prepared. You have equipped yourself with a pill of your own, recently placed on the market - let us call it an ED:P2 - one which has the following two effects. First, it produces a mark on your forehead - a mark that can only be produced by taking just such a pill - and second, it hardwires you to follow through on any threat you might have made or will make within, say, an hour either side of having taken the pill. Now then, would it be rational for you to take your pill if I had already taken mine? Clearly not. Would it be rational for me to take mine if you had already taken yours? A fortiori not. So the winner, it would seem - no less in the game of Extortion than in the one called Domestic Dispute - is whoever manages to, at it were, draw first on his or her irrationality holster!

As the need arises we will be looking at yet other games. For now, however, suffice it to say that what, among other things, all these games have in common, is that each is designed to match a dilemmatic pattern of human interactivity with a disposition remarkably well suited to doing the best one can under these specific dilemmatic circumstances, coupled, of course, with the unveiled suggestion that these dispositions are, in fact, nothing more (nor less) than rational, strategic, intramental responses to these patterns. Moreover, we claim, that a disposition be accounted for in just this way is a necessary, albeit insufficient, condition of its being a moral disposition. So, first: if a disposition is not to be accounted for in this way, then it cannot be a moral disposition. (The implications of this view, dire as they may be, will be explored later.) And second: what further conditions must be satisfied (esoteric though they may be) in order for a disposition to count as moral will be explored later still.

15. The Logic of Precommitment in General:

But here is something else all these games have in common. In each case the (suggested) response thereto involves a ploy game-theoreticians call a 'precommitment' strategy. That is, in each case the solution to the dilemma is to do something either in the world - e.g. in the DD transparently putting in place a Doomsday Device - or else to the mind - e.g. in the PD and ED transparently taking a pill - the upshot of which being one's thereby being precluded from doing what she would otherwise be rationally inclined to do should a certain situation arise. And the rationale for doing so is that by doing so she hopes to preclude the situation - under which she would otherwise be rationally inclined to do what she is now precluded from doing - from arising in the first place.

But having already divided precommitment between external and internal techniques, we can multiply our options further by noting that some precommitments are designed to preclude oneself from being the author of one's own demise - e.g. Ulysses ordering his men to tie him to the mast - whereas others are adopted to preclude the co-player from precipitating the unwanted state of affairs - e.g. in the DD dissuading the Soviets from first-striking. Let us capture this distinction with the terms 'self-disciplining' and 'other-disciplining'.

And we can multiply our options even further by noting that some precommitments leave the agent preferentially well-ordered, a.k.a. rationally 'content' - e.g. in the DD, where upon provocation the American President actually wants to retaliate - whereas others render the agent 'epiphenomenally frustrated' - e.g. in the Hobbesean solution to the PD where the subject wants to disobey the Sovereign but cannot.

It would seem, therefore, we have available to us no fewer than eight sub-species of precommitment strategies. These are

1) External Self-disciplining epiphenomenally Frustrated Precommitment, or ESFP,
2) External Other-disciplining epiphenomenally Frustrated Precommitment, or EOFP,
3) External Self-disciplining Contented Precommitment, or ESCP,
4) External Other-disciplining Contented Precommitment,

or EOCP,

5) Internal Other-disciplining Contented Precommitment, or IOCP,
6) Internal Self-disciplining Contented Precommitment, or ISCP,
7) Internal Other-disciplining epiphenomenally Frustrated Precommitment, or IOFP, and
8) Internal Self-disciplining epiphenomally Frustrated Precommitment, or ISFP.

Now let us take a moment to familiarize ourselves with these options and to note the merits and demerits of each.

The classic case of ESFP is, as already r[illegible]ed, Ulysses ordering his men to tie him to the mast. Here is another:

> Much as I love being up early in the morning, like most people I loathe getting up. And so I authorize (whomever) to drag me bodily from my bed, if need be kicking and screaming.

Note that here the experience of the violation of my autonomy is acceptably short-lived. A more problematic case would be:

> Much as I know I need to save for retirement, I know too there will always be what appear to be more pressing demands. And so I irrevocably authorize the pay-roll office to make pension deductions from my salary.

Here my epiphenomenal frustration is both deep and on-going. But at least I can anticipate a point at which I will be glad of having made these sacrifices. But a more problematic case yet would be:

> Much as I know I ought to tithe, I know too there will always appear to be more pressing demands. So I irrevocably authorize both the payroll office to make charitable deductions from my salary and the pension office to make such deductions in the future from my pension.

For here, it might be argued, my rationale borders on the incoherent. For if, in the absence of an external pre-commitment mechanism I have shown myself to be consistently unwilling to tithe, then what sense can be made of my claim I know I ought to?

We have already seen that the classic case of EOFP is solving the DD by means of putting in place a so-called Doomsday Device. Once my sensors detect provocation my own silos retaliate automatically. Here, as we have seen, this epiphenomenal frustration produces profound ethical and legal ambiguities. But we need not confine ourselves, as does Kavka, to our intuitions regarding nuclear deterrence. Here is an equally instructive case in point:

Am I responsible for the mangling of an intruder even though I have clearly posted the warning, "Pit Bull on Premises!"? Or is he the author of his own misfortune? Clearly I would have been culpable - probably for the use of excessive force - had I instead been home and set the dog on him. That is, I would have been culpable had I not precommitted. Can I be allowed to escape culpability in virtue of nothing more than being able to claim that, having precommitted, the matter was out of my hands? But, I might remind the court, the dog's attack was conditional upon the intrusion; moreover this conditionality was designed to discourage the intrusion. But then, counters his solicitor, of what relevance then is the precommitment? If it was permissible to set the trap it would have been equally permissible to spring it myself in the event of intrusion. Since I have conceded that springing the trap myself would have constituted excessive force, likewise must I concede that setting the trap was excessive.

Thus, it might be argued, since Kavka would be wrong to reject the Wrongful External Precommitment Principle (or WEPP), i.e.

> if it would be wrong to do something under certain conditions, then it would be wrong to employ an external precommitment mechanism to do that thing should those conditions arise,

likewise is he wrong to reject the Wrongful Intentions Principle (WIP), i.e.

> if it would be wrong to do something under certain conditions, then it is wrong to form the intention to do that thing should those conditions arise,

since, that is, the WIP is a.k.a. the Wrongful Internal Precommitment Principle (or WIPP). For how can it be right to precommit internally - whether with 'content' or 'epiphenomenal frustration' - with the intent of disciplining my co-player, and yet wrong to precommit externally - whether with 'content' or 'epiphenomenal frustration' - with the same intent?

Or at least that is one intuition. Here is the other. Since it is manifestly right to precommit externally - whether with 'content' or 'epiphenomenal frustration' - with the intent of disciplining my co-player, and since no moral distinction between the two strategies can be found, it follows it is likewise right to precommit internally - whether with 'content' or 'epiphenomenal frustration' - with the same intent. The WIPP and WEPP do indeed stand or fall together. But, as it happens, they stand rather than fall!

And yet a third intuition is that the WIPP and the WEPP stand or fall severally. Might this intuition be a cousin to the commission/omission distinction? And might both intuitions have a common relation in the conviction that there is relevance in the proximal/distal distinction? Hard to say. Easier to say, however, is this. To suppose these conflicting intuitions can be readily resolved is to suppose the entire corpus of jurisprudence on responsibility has been an industry built on a mere chimera. Safer, I submit, to move on.

Now it might be supposed that the very idea of an EOCP strategy is, if not incoherent, then at least unmotivated. For whereas an EOFP strategy could be warranted, as we have already seen, by my being transparently inclined in the absence of an external mechanism to do that which, were I expected to do so, the expectation of my doing so would encourage my co-player to behave contrary to my druthers, if I were instead contented with my consequent disposition, then, it would seem, there would be no need to externally enforce it. But if the incoherence is between the externality and the epiphenomenal frustration, then likewise in-coherent should be any ESCP strategy. But then what about the following case?

> Much as I love being up early in the morning, like most people I loathe getting up. And so I authorize (whomever) to lift me, still sleeping, from my bed, and to put me bodily into the shower.

Note that since I am asleep at the moment of my being 'interfered' with, at no point do I experience a violation of my autonomy. So for EOCP, why not the following?

> Much as I love being up in early in the morning, like most people I loathe getting up. But much as I love being up and loathe getting up, I loathe even more my wife's staying in bed. But she will stay in bed only if I do. So I authorize (whomever) to

> lift me, still sleeping, from my bed, and to put me bodily into the shower.

So cases of EOCP seem possible after all.

Now as already noted the classic case of IOCP is the American President making himself into a Russo-phobe. And so a typical case of ISCP would be my making myself into a miser (in the retirement case), or (in the tithing case) making myself less equivocally philanthropic. A case of IOFP, in turn, is Gauthier's solution to the PD. For note that, for all that has been said so far about the CM disposition, upon recognition of a fellow CMer the CMer need not actually prefer cooperation over exploitation. Nor does she cooperate only because she knows she cannot exploit - 'cannot' in the modal sense, i.e. there is no possible world in which she defects while the other cooperates. (More of this later.) Rather she cooperates because she cannot defect - 'cannot', that is, in the sense that she is incapable of defection. But this 'incapacitation' can be, and often is, 'epiphenomenally frustrating' in the extreme! (As I say, more of this anon.)

What remains, then, is ISFP, the paradigm case being my imposing a duty on myself to save or tithe - i.e. pangs of guilt if I default - while all the while bridling under the weight of my self-imposed duty. But note that it is only in these cases of ISFP and, as we have already seen, of ESFP, that the issue of coherence seems to arise. For it is one thing to observe that people can (and do) employ both external and internal mechanisms to force themselves to do what they would rather not, and that they do seem to be able to live with such epiphenomenal frustration. It is quite another to concede that putting themselves in such a position is always, or even sometimes, rational.

But if the locus of the incoherence is in the experience of the tension, i.e. in ISFP and IOFP, then this incoherence should likewise be posited of ESFP and EOFP, since therein likewise will be found the experience of this tension. In which case we should find all of IOCP, ISCP, EOCP and ESCP equally unproblematic. But if we judged all cases of content to be equally unproblematic, how do we explain our discontent with G-D and G-D-G? Goldman's becoming G-D was a case of ISCP, G-D's becoming G-D-G a case of IOCP. So now it seems the problem lies less in the content/frustration distinction than in ... in what? The internal/external distinction? But surely that cannot be relevant. The difference between self-disciplining and other-disciplining? Why should that matter? Or is there some fourth distinction yet to be discovered?

Indeed there is. For what we have yet to distinguish between are precommitment strategies - be they E or I, O or S, C or F - which need be done 'transparently', and those which can only be effective if opaque, or 'cloaked'. That is, like a quarterback who calls a passing play, falls back and then cocks his arm in such a way that he could be readying himself to pass _or_ hand off, cloaked precommitments work precisely because they do _not_ give themselves away.

But, while we are at it, let us distinguish too between precommitment strategies by which I cloak my own precommitment from you, and those - odd though it may at first seem - by which I blind _myself to yours_. And - as if that were not odd enough - let us distinguish as well between blinding myself to _your_ precommitment and blinding myself to my _own_!

The importance of all this, especially for understanding Goldman and Goldman-turned-Dorff, will emerge in due course. For now suffice it to observe that with the transparent/cloaking distinction we have not eight species of precommitment strategies, but sixteen. So let us symbolize _this_ multiplication by prefixing each of those we already have with either a 'T' or a 'C'. We can capture the notion of blinding oneself to the precommitments of others, a.k.a. 'self-blinding', with a 'B'. And let us symbolize the blinding of oneself to her own precommitments, a.k.a. 'self-self-blinding', with an 'SB'.

We have had plenty of examples of transparent pre-commitment strategies already. And we will encounter clear cases of cloaking, self-blinding, and self-self-blinding in the next chapter, when we look at strategies for winning at the game of Chicken. But in the meantime for a case of self-blinding we need look no further than a re-look at the game of extortion. Recall that taking the irrationality pill in the case of the parent, i.e. the ED:P1, or the threat-enforcer pill in the case of the extortionist, i.e. the ED:P2, will work if but only if the pill is taken transparently; if, that is, it announces itself by its tell-tale mark on the forehead. But suppose there is a pill on the market - call it ED:P3 - which produces the following two effects. First, it produces a particular mark on one's forehead, a mark that can only be produced by ingesting this particular pill. And second, it renders her blind to whatever marks - or at least to the threat enforcer mark - that might appear on the extortionist's forehead. Would it be rational for the extortionist to take his ED:P2 if the parent had already taken her ED:P3, even if she had yet to take my ED:P1? Clearly not. So, it would seem, there are games the winning strategy for which may involve a transparent precommitment not to be transparently precommitted on; that is, a precommitment not to look at whatever the co-player may have transparently precommitted to. So whereas we might normally suppose it would be a marvelous advantage to be able to read other people's minds, sometimes, it would seem, it can be a decisive advantage not to be able to do so. But, as I say, more of this in due course.

16. Precommitment and Decidability:

In the meantime we are now in a position to suggest (at least) a (line of) defense for Gauthier against what would otherwise count as a subtle but devastating attack against his entire program.

The general form of the objection is this. Recall that Gauthier's project is inter-level reduction; and that a condition of its success, therefore, is that at no point can it appeal to an explanatory or justificatory principle which is itself a moral one. So if it can be shown that such a principle is being appealed to at any juncture - no matter how subtlely or veiled - then the project not only fails, it fails in its entirely. Here I confine myself to only one such argument, recently advanced by Richmond Campbell. [55] But if the defense works against Campbell, perhaps it can be generalized and applied in Gauthier's defense against other objectors. Campbell's argument is as follows:

> So far we have been confining ourselves to PDs in which the fruits of mutual cooperation are divisible but not divisibly divisible. That is, if both Column and Row keep the faith they will each serve two years, as opposed, recall, to three had they both ratted. Note, however, that though they have just gained a total of twenty-four months of freedom between them, they are not free to divvy up this benefit as they themselves see fit. In fact an egalitarian distribution of the gain is being imposed upon them by the prosecutor. In most situations of cooperation, however, the parties to it are free to distribute its dividends. And this fact imposes an addition hurdle to their cooperating in the first place. For example:
>
> Suppose severally Column can produce two widgets a week and Row can produce four, but cooperatively they could produce ten. Discounting, for simplicity, any considerations of utility marginalization or other externalities, there are five distribution schemata, each

of which is Pareto-optimal. A distribution scheme is said to be Pareto-optimal just in case

> there exists no alternative distribution scheme such that, relative to the scheme in question, someone could be made better off without someone being made worse off.

Let us assume that no two parties would cooperate unless both were assured of a distribution scheme which offered at least some improvement to her net take, and that they would then settle at a split that was, at the very least, Pareto-optimal. So, it would seem, Column can improve her position by one, two, or three widgets a week, and Row can improve hers by three, two, or one, respectively. But which of these three schemata will it be?

Since we are assuming non-tuism Column should be able to reason, quite rightly, that Row would be a fool to refuse to cooperate even if all Column offered her was a net gain of only one widget. But similarly can reason Row. Thus clearly a condition of extricating themselves from their PD is that Column and Row agree to some principle beyond mere strong Pareto-optimality by which they will distribute the profits of their cooperation.

Now some thinkers argue - among them Kai Nielsen [56] and, albeit for different reasons, John Rawls [57] - that Row should content herself with an egalitarian five and five, i.e. a net gain for herself of only one. Others - among them Robert Nozick [58] and, albeit for different reasons, Jan Narveson [59] - opine that Row is entitled to seven and Column only three. And Gauthier, opting for an algorithm called 'minimax relative concession', suggests they settle at four for Column and for Row six. [60] But whereas Nielsen, Rawls, Nozick and Narveson are all prepared to concede - or at least they can be forced to concede - that their algorithms are informed by intuitions which are unabashedly moral, Gauthier must be able to show that minimax relative concession is in no wise so informed. And this, claims Campbell, he neither has shown in Morals by Agreement, nor can he show it. Thus, concludes Campbell, insofar as

1) the Bargaining Dilemma (or BD) cannot be resolved without recourse to a moral principle, and insofar as
2) none but the least interesting PDs can be resolved without a resolution to the BD, it follows that

3) none but the least interesting PDs can be resolved without recourse to a moral principle, and hence that

4) Gauthier's reduction - of the moral to the non-moral, the pre-moral, the amoral, call it what you will - fails.

But, on Gauthier's behalf, let us take a closer look at the BD. What Campbell's analysis of the BD overlooks, it seems to me, is that both Column and Row are as keenly committed to resolving their impasse (a.k.a. broad compliance) as they are to doing as well as the logic of the situation might allow them (narrow). [61] But since they have no access to any principles, moral or otherwise, other than those already on the table, they might, indeed they would, I submit, submit to the following reasoning:

What is preventing Column and Row from striking a bargain, it seems, is each is availing herself of a first-order transparent precommitment pill. Call it the BD:P1. That is, realizing that in the absence of taking such a (narrow compliance) pill she would settle for as little as one extra widget, Column has transparently precommitted herself not to settle for less than two. Unfortunately, Row has likewise transparently precommitted herself, and so no resolution is possible. But suppose there were a second-order pill on the market - a BD:P2 - one which produces the following two effects. First, it produces a mark on one's forehead that can only be produced by taking just such a pill. And second, it hardwires her not to take the BD:P1 when bargaining with those, but only those, with just such a mark on their foreheads. Would it be rational to take such a BD:P2? Clearly it would. Would it resolve the BD? Well, perhaps not yet. But then suppose there were yet a third pill on the market - a BD:P3 - and so on. Is there an nth-ordered pill it would be rational for both Column and Row to take which, if taken by both, would resolve the BD? Gauthier's answer, it seems, is: yes. And, he suspects, it is such that it would produce an outcome co-extensive with that produced by a mutual commitment to minimax relative concession.

That said, I leave it for another day - and perhaps too to a more committed defender of Gauthier - to actually show there is a pill that would resolve the BD, and that co-extension with minimax relative concession is what it would produce. But if such a proof can be generated - and I suspect it can - then it would follow that Gauthier can solve the BD, and hence the PD, without recourse to an undischarged moral principle.

4. PRISONERS' DILEMMA IS NOT A NEWCOMB'S PROBLEM [1]

1. Background:

Among contemporary contractarians, internalism of one sort or another has been, for some time now, virtually the only game in town. So, I predict, it will remain. Some years ago, however, David Lewis pointed out that, while preserving all relevant formal features, the payoff structure in another dilemma, namely Newcomb's Problem (NP), can be re-rendered as identical to that in a PD. [2] An NP, baldly put, is this:

Placed before me, by God let us say, are two boxes, one transparent, the other opaque. The transparent box contains a thousand dollars, the opaque box, I am told, either an additional million or nothing at all. And, I am told, I can have either the opaque box or both. But here is the catch. If God predicts I will take the thousand as well as the possible million, She will not have put the million in the opaque box. If, but only if, She predicts I will forgo the thousand, then She will. [3]

Now from the fact that the payoff structures in the two dilemmas are, as we will see, identical, Lewis concludes - all too quickly, I shall argue - the two dilemmas are, for all intents and purposes, one and the same. Since Lewis does not think we can simply think our way out of, i.e. 'deconstruct', an NP, he stops short of concluding we need look no further for a solution to the PD - i.e. to Hobbes and/or beyond to Gauthier. More recently, however, John Leslie has resurrected Lewis' argument for the 'equivalence' of the two dilemmas and concluded instead that, since we can think our way out of an NP, likewise can we think our way out of a PD. [4] If Leslie is right it would follow we can forgo Hobbesian and/or Gauthierian solutions to the PD. And if we can forgo Gauthier's solution to the PD, then, I am loathe to concede, my own project is dead in the water!

In what follows, then, I do four things. First, I argue with Leslie and against Lewis that we can think our way out of an NP. So if Lewis and Leslie were right about the equivalence thesis Leslie would be right about the deconstruction thesis. Next, however, contra Lewis and Leslie I show why the equivalence thesis is false. Having thus shown that the PD cannot inherit its deconstructability from the NP, I show third that neither can the PD be deconstructed in its own right. And, finally, insofar as I will be arguing that Lewis' and Leslie's mistake lies in failing to properly characterize the independence condition, towards the end of this chapter I want to draw attention to the correct role the independence condition plays in defining the PD, and hence in defining our pre-, non-, or a-moral predicament.

2. Equivalence:

Because Lewis' case for the equivalence thesis is as concise as it is ingenious, I will quote it here almost in full:

> You and I, the 'prisoners', are separated. Each is offered the choice: to rat or not to rat.

"Ratting," explains Lewis, "is done as follows:

> One reaches out and takes a transparent box, which is seen to contain a thousand dollars. A prisoner who rats gets to keep the thousand. If either player declines to rat, he is not at all rewarded; but his partner is presented with a million dollars, nicely packed in an opaque box. Each faces a long sentence and a short sentence to be served consecutively; escape from the long sentence costs a million, and escape from the short sentence costs a thousand.

But, Lewis reminds us,

> it is irrelevant how [we] propose to spend [our] money. So the payoff matrix looks like this. If we both rat, we each get a thousand [5],

with which, presumably, we will each buy off our short (i.e. two-year) sentences, leaving ourselves each with three. If neither of us rats, we each get a million, with which, presumably, we will each buy off our long (i.e. three-year) sentences, leaving ourselves each with two.

But if one of us rats while the other keeps the faith, then the ratter gets both the thousand and the million - and so can afford to walk - while his partner gets neither - and so languishes the full five. "There we have it," concludes Lewis, "a perfectly typical case of Prisoners' Dilemma." [6]

Now then, he continues, "[m]y decision problem, in a nutshell, is as follows; [and] yours is exactly similar.

> 1) I am offered a thousand - take it or leave it.
> 2) Perhaps also I will be given a million; but whether I will or not is causally independent of what I do now. Nothing I can do now will have any effect on whether or not I get my million.
> 3) I will get my million if and only if you do not take your thousand.

Newcomb's Problem, claims Lewis,

> is the same as regards points (1) and (2). The only difference - if such it be - is that point (3) is replaced by
>
> 3') I will get my million if and only if it is predicted that I do not take my thousand,

which, as Lewis points out, reduces to

> 3") I will get my million if and only if a certain potentially predictive process yields the outcome which could warrant a prediction that I do not take my thousand.

But, of course, Lewis continues,

> t]he potentially predictive process par excellence is simulation [or replication]. [7] ... [So a] special case of (3") [is]
>
> 3"') I will get my million if and only if my replica does not take his thousand. [8]

But, claims he, it is equally clear that

> t]he most readily available sort of replica of me is simply another person, placed in a replica of my pre-dicament. For instance: you, my fellow prisoner ... [Thus a] special case of (3"') [is] (3)! Inessential trappings aside [then], Prisoners' Dilemma is a version of Newcomb's Problem, quod erat demonstrandum. [9]

3. Equivalence and Deconstruction:

Q.E.D. indeed! agrees Leslie. [10] But whereas Lewis and Leslie agree about the two dilemmas being equivalent, they disagree about their deconstructability. By Lewis' lights,

> [just as] it is rational [in a] Newcomb's Problem [for me] to take the thousand no matter how reliable the prediction that you will [do likewise] may be - the reason being that one thereby gets a thousand more than he would have if he declined, since he would get his million or not regardless of whether he took his thousand - [so] it is rational [in a] Prisoners' Dilemma [for me] to rat no matter how alike [we] may be, and no matter how certain [I] may be that [you] will decide like[wise] - the reason being that one is better off if he rats than he would be if he didn't, since he would be ratted on or not regardless of whether he ratted. [11]

Leslie, on the other hand, counts himself among those who, in Lewis' words, think that

> [just as] it is rational [in a] Newcomb's Problem [for me] to decline the thousand if the predictive process is reliable enough - the reason being that those who decline their thousands will probably get their millions - [so] it is rational [in a] Prisoners' Dilemma [for me] not to rat if [we] are enough alike - the reason being that those who do

> not will probably not be ratted on by their like-thinking partners. [12]

How is it, then, that Leslie's rationality allows him to deconstruct - i.e. to think his way out of - an NP/PD, while Lewis' rationality does not afford him the same facility?

4. Deconstructionism, a.k.a. the Symmetry Argument:

Leslie's argument runs as follows. If

1) I have good grounds to believe - and, thinks he, the very nature of the PD ensures I do - you and I are virtual replicas of each other - that is, we are identical with respect to the inputs and algorithms involved in the choice situation at hand [13] - then
2) I likewise have good grounds to believe that on your side of the wall you will be doing precisely what I will be doing on mine. [14] Thus,
3) in taking the thousand in an NP - or, *mutatis mutandis*, in a PD ratting - I assure myself of your taking the thousand - ratting. Likewise, *mutatis mutandis*, to assure myself of your forgoing the thousand - keeping the faith - all I have to do is forgo the thousand - keep the faith - myself. [15]

But, reasons Leslie, if this is right then, as it turns out, there never were four possible action-combinations after all. And what is more, I can know this. Our only options are, and always were, for us both either to take the thousand - rat - or else both forgo the thousand - keep the faith. And since

4) I fare better by our both forgoing the thousand - keeping the faith - than I do by our both taking the thousand - ratting - it follows quite straightforwardly that

5) I ought to forgo the thousand - keep the faith.

Quod erat demonstrandum. The NP/PD is and always was a problem/dilemma in name only.

Now then, unless Plato, Hobbes and Gauthier have been tilting at windmills these last two and a half millenia, something has gone terribly amiss. Either Leslie is cheating on Lewis or else Leslie _and_ Lewis are cheating on the rest of us. But which is it?

This much is certain. If Leslie is mistaken in his argument he is in excellent company. Lawrence Davis [16] for example - following the lead of both Anatol Rapoport [17] and John Watkins [18] - likewise argues that,

> If we [can] assume that each prisoner knows that each knows that each is rational, as well as

> knowing the information represented in the matrix, it appears that Row is not uncertain what Column will do. More precisely, he can and will determine what Column and he himself should, and so by assumption will, do. He knows that Column is a rational agent and that he himself is. He knows further that their situations and information are symmetric: whatever considerations would lead Row himself to choose one way would lead Column to choose exactly the same way. The information Row has, then, implies that the outcomes <cooperate, defect> and <defect, cooperate> are not in fact possible. Of the two outcomes remaining, <cooperate, cooperate> is rated higher by both Row and Column. The rationally prescribed alternative for each, then, is to [cooperate]. And, by assumption, each will take this alternative, secure in the knowledge that the other will take it also. [19]

However in response to an earlier version of this argument proffered by Watkins, Amatya Sen points out that

> [i]t is precisely because [Column's] choice cannot be assumed by [Row] to be a mirror-reflection of his own choice that the dilemma of the prisoners is supposed to arise ... [And yet] Watkins essentially makes each prisoner assume that the other prisoner's action will be a function of his own action and will in fact coincide with it. [20]

To which Davis replies that "the choices and actions of each [prisoner can] be a function of what is rationally prescribed for each" without, on that account, being a function of those actual choices and actions. "Cf.," says he,

> two clocks known to work perfectly and to have been synchronized at an earlier time. By looking at one we can know what time the other indicates, which is not to say we can make the other one indicate an arbitrarily selected time by setting the first to that time. [21]

What Davis calls "synchronicity" here Leslie dubs "quasi-causation". [22] But their points are essentially the same. Column might be tempted to think, at the very moment she (and therefore Row too) is about to cooperate, at the last minute she, Column, will defect. But ex hypothesi she will then remember that Row, being no less clever than Column, will likewise think to defect at the last minute. So the only way for Column to ensure Row will cooperate is to commit herself - and therefore, thinks Leslie, Row along with her - not to defect at the last minute. [23]

5. Why the NP Is Deconstructable:

With respect to the deconstruction of the NP I think Leslie et al are right. My reasoning is as follows:

If I know - or even have good grounds to suspect - you and I will be thinking our way identically through our predicament, then it is irrelevant that there is a fact of the matter about which action you have ultimately settled on. True, that fact of the matter is fixed. And so I might be tempted to think, along with Lewis, that whichever action you have taken I might just as well take the thousand - or, mutatis mutandis, rat. But what does it mean to say "you and I will be" - or are even likely to be - "thinking our way identically through our predicament"? On Lewis' own terms - he is, after all, the modal realist par excellence [24] - it means:

1) of all the possible worlds accessible to this one, none - or at least damn few - are such that you and I ultimately settle on different strategies.

True, in some of these worlds - say, half - we each take the thousand - rat. So whereas I am, in a very real sense, holding my breath to see of which of these two kinds of worlds the actual world turns out to be, I need not be overly concerned it will turn out to be one of the rare - if not non-existent - ones in which we ultimately settle on different strategies. Now then, assuming, as I must,

2) what strategy I ultimately settle on in some sense, at least, determines of which of these two world-kinds the actual world will turn out to be, and (just as trivially) that
3) there is at least some sense in which I have a choice over which strategy I ultimately settle on, and since
4) Leslie's (4) (above) holds unproblematically,

5) likewise does his (5).

So Leslie is not cheating on Lewis. Rather even on his own terms Lewis is cheating himself out of a solution to the NP.

6. Equivalence and Independence:

If Leslie and I are right about the deconstructability of the NP, and if Lewis and I are right about the non-deconstructability of the PD, then Lewis and Leslie must be wrong about the equivalence thesis. So wherein lies the cheat?

It lies, I submit, in Lewis' claim - apparently acceded to by Leslie - that "Newcomb's Problem is the same as regards points (1) and (2)." More specifically it lies in the claim that the NP and the PD are both characterized by

2) Whether I will also be given a million is causally independent of what I do now.

For whereas (2) is true of the PD, of the NP, I claim, it is patently false.

That "whether I will also be given a million is" not "causally independent of what I do now" can be made clear, I submit, by redescribing the NP as follows:

> Before me are two boxes, both transparent, one containing a million dollars, the other a thousand. The box containing the million, however, is bottomless, and it is placed firmly over a trap door, and this trap door, in turn, attached to the lid of the box, such that if I open the box to retrieve the million the trap door will open and the million will fall back down to God from whence, ex hypothesi, it came in the first place. Or, in what Lewis presents as the PD ver-sion of the NP, the million will fall down into the lap of my co-player. The aforementioned 'attachment', however, is in turn attached to the box containing the thousand, which is poised in perfect equilibrium on a fulcrum the other side of which is a steep shute which likewise leads back down to God from which it came in the first place. Or, in what Lewis presents as the PD version of the NP, the thousand will slide down into the lap of my co-player. Thus by pushing the box containing the thousand down the shute and hence irretrievably out of reach - and

> only by doing so - can I snap the first attachment, thereby disabling the trap door, thereby enabling me to safely open the box containing the million, thereby enabling me to retrieve the million.

Now then, is it true that "whether or not I will be given the million is causally independent of what I do now"? Clearly it is not. Then in virtue of what in the original NP do I claim my redescription of it is accurate? In virtue of the ex hypothesi retro-causal powers of God's foreknowledge!

But, it may be objected, how can mere knowledge be a causal power? To which I answer, How could it not? Suppose in the PD there is a one-way glass dividing the two interrogation rooms. Or perhaps a listening device. Suppose Column can see or hear Row, but Row cannot see or hear Column. Could it be said Column's behaviour remains causally independent of Row's? Clearly not. But then if Column's behaviour is no longer causally independent of Row's, in what sense is God's behaviour - i.e. Her putting or failing to put the million in the opaque box - causally independent of mine, i.e. taking or forgoing the thousand?

Furthermore, under the circumstances just described, could anyone conscionably advise Column to keep the faith if Row had just ratted? Clearly not. Could anyone conscionably advise Column to keep the faith if Row had just kept the faith? _A fortiori_ not. So since the conscionable thing to counsel in the NP is to forgo the thousand, whereas the conscionable thing to do in the PD _under identical epistemic-turned-causal conditions_ is to rat, it follows that the NP and the PD cannot be equivalent. _Quod erat demonstrandum_.

7. _Some Side-Issues_:

Furthermore, by re-rendering the NP as we have, we are now in a position to resolve two further issues that divide students of the NP. The first has to do with the coherence of the scenario once divine anticipation is allowed into it. And the second has to do with whether it makes a difference whether the predictive process is apodeictic or just 'pretty damn good'. For we can now see that the NP does _not_ depend in any important way on anything metaphysically suspect, like divine anticipation. For we can as easily imagine a mad but insightful game-theorist setting things up as described above in order to conduct empirical tests into the nature of human rationality. And as to the issue of apodeicticity versus mere reliability, we can _see_ whether this makes a difference by imagining further devices attached to the aforementioned 'attachments', such that, for example, by pushing the box containing the thousand down the shute, there is only an 83% probability this will snap the attachment preventing me from retrieving the million. There might be some residual dispute as to which model of rationality would be most appropriate, e.g. straightforward expected utility maximization, minimax, disaster avoidance, or what have you. [25] But it is clear that moving from apodeicticity to mere probability does not of itself suggest the problem is now no longer deconstructable.

8. Can the PD be Deconstructed in Its Own Right?

But, it will be objected, what I have shown so far, and all I have so far shown, is that the causal independence condition - which, claim the symmetrists, is definitive of the PD and NP irrespectively - can be defeated by epistemic access. But, the symmetrist might argue, she can repair her position by

1) supplementing the causal independence condition with an epistemic cloaking condition,

2) insisting that this supplementary condition nonetheless applies to both the NP and PD irrespectively, and then
3) claiming that notwithstanding this cloaking we can nonetheless think our way out of this cloaking and we can do so in the NP and the PD irrespectively.

For what takes the place in the PD of my knowledge of God's foreknowledge in the NP is my knowledge, as Davis puts it, that

> each prisoner knows that each knows that each is rational, as well as knowing the information represented in the matrix.

So, claims the symmetrist, since I have already conceded we can think our way out of the NP, the onus remains on me to show we cannot do likewise with respect to the PD. Or, to put her challenge another way: even if I have shown the two dilemmas are not equivalent - and therefore the PD cannot *inherit* its deconstructability from the NP - it by no means follows the PD cannot be deconstructed in its own right. And that, I concede, *does* remain to be shown. What remains to be shown more specifically is that, whereas in the NP neither can Column *hope* Row will cooperate while Column defects, nor need Column *fear* Row will defect while Column cooperates, in the PD, by contrast, grounds for such hope and fear remain.

To show this, I will advance my case in three stages. First I export the symmetrist's challenge to yet another dilemmatic game, namely that of Chicken. Next I argue for the non-deconstructability of Chicken. And third I show that, in those respects germaine to the issue at hand, the PD is equivalent to Chicken. I begin, then, with the rule-book for Chicken.

9. How to Play Chicken:

The game of Chicken, henceforth the CD, has a long and honourable history in the animal kingdom - Walt Disney made a genre of it - as well as a long and, some might say, dishonourable history in human affairs, both interpersonal and international. But perhaps the most celebrated case in point is the testosterone ritual developed in small-town America in the late Fifties and early Sixties. Late on a Saturday night the unmarried males and females of the species would repair to a relatively isolated strip of blacktop a few miles from town, preferably straight but unshouldered. Each of two contestants - once again we will call them Column and Row - would proceed in opposite directions to his starting position about a mile from the theoretical point of impact, turn around, wait for the signal, and then -

floor it! The first to chicken out was then understood to have lost not only his oil pan (approximately a week's salary) and his place in the pecking order among his peers (a.k.a. face), but, more importantly - this being why the attendance of the females of the species was required - sexual prowess. And the one who by virtue of the other chickening out first does not himself have to chicken out was understood to have gained face among his peers and, presumably, prowess in the adoring(?) eyes of the female onlookers.

So the choice matrix looks like this. Each prefers that the other chickens out first so he does not have to. Failing that he prefers they both chicken out. (The reason for this is that face is, so to speak, a zero-sum utile.) Failing that he is prepared to chicken out himself. But on no account does he want neither of them to chicken out since - loathe as he is to lose oil pan, face and prowess - he would rather lose all three than lose his entire car, the rest of his head, and - not to put too fine a point on it - the physical manifestation of his manhood.

Note that in both the PD and the CD Column's and Row's preferences coincide twice; but whereas in the PD these are their second and third choices, in the CD they are their second and fourth. Furthermore, in a PD defection dominates; that is, irrespective of whether Column and Row expect Row and Column respectively to defect or cooperate, Column and Row respectively would be well advised to defect. But in a CD, by contrast, neither chickening out nor staying the course dominates. If Row chickens out then Column should stay the course. But if Row stays the course then Column should chicken out. In both games the choices are mutually exclusive and exhaustive; and in neither game is the non-tuism condition violated. But the independence condition, so definitive of the PD, is precisely what is missing in the CD. Or at least it is only three-quarters satisfied. That is, by chickening out Column can cause Row to stay the course. But by staying the course Column can only encourage Row to chicken out.

Just how strongly and for how long Column's staying the course will encourage Row to chicken out will depend, of course, on the strength of Row's resolve to stay the course. But the strength of Row's resolve, in turn, will depend on the strength of Column's. And so on in a logically infinite - albeit temporally finite - loop. But the only means by which Column and Row can communicate the strength of their respective resolves is, unfortunately, too course-grained. For whereas the precise point at which the loser chickens out precisifies the weakness of his resolve, the winner's having stayed the course indicates only that the point at which he would have chickened out - which remains to be determined - has yet to be reached. Thus since Column can always hope that Row's resolve will falter before his own, he can always rationalize himself into staying the course. But, of course, likewise can Row rationalize himself into staying the course. And this is why chickening out, if it takes place at all, takes place only a fraction of a second before the (sometimes mis-)estimated point of impact.

That being the case, we can now see that the rationale behind opting for a straightaway is not so much to provide each player an opportunity to test and show his mettle, but rather and simply to allow the contestants to reach lethal speed and provide suspense for the spectators. But, it would seem, both of these objectives could as readily be served by placing the theoretical point of impact on a blind curve. That is, once we see, as we soon shall, that winning in a CD amounts to being one conditional precommitment step ahead of the opposition, it makes no difference whether this step is taken en route on the straightaway, en route to the blind curve, at the start of the straightaway, or at the start of the blind curve. All we really need is that there be at least a logical moment - as distinct from a temporally extended one - in which each players can process the fact that his co-player has yet to chicken out.

This is important. For it might be imagined that a further distinction between the PD and the CD is that in the latter, but not the former, timing is of the essence. That is, in the PD it is irrelevant to Column's choice situation whether Row's cooperation or defection was executed yesterday, today, or tomorrow. (In fact in the PD even knowing both the when and the what of Row's doing can in no wise effect Column's choice since, as we have already seen, the rational thing for Column to do in any case is defect.) In the CD, by contrast, Column's and Row's moves and countermoves are supposed to be intervisible. And, as I say, it might be thought this is significant. But, as we are about to see, we can always construct our CD's in such a way that the logic of the dilemma, though sensitive to logical antecedence and consequence, is indifferent to temporal priority and posteriority. That is, all these moves and countermoves could as readily be preprogrammed days before the event, and so programmed in mere anticipation of what countermove the counter-anticipation of each move would inspire. But if each player's moves and countermoves are so preprogrammed, they would not, on that account alone, be any the less moves and countermoves in a CD.

For example it might be supposed that if they have decided to play on a blind curve the situation has altered, because now Column can delay precommitting himself. But, one might well ask, to what avail? Suppose he chickens out before the point of intervisibility. Then why did he pretend to be willing to play in the first place? In the hopes Row will have chickened out too? But then is this not just the game of Proto-Chicken?

The game of Proto-Chicken is the same as the game of Chicken, save that

1) winning is done by being 'macho' enough to play real Chicken while the other proves himself too 'chicken',
2) neither-winning-nor-losing-so-badly is done by being too chicken to play Chicken while the other proves too chicken too,
3) losing is done by being too chicken to play Chicken while the other, by contrast, proves himself macho enough to play, and

4) losing-big is done by being macho enough to play while the other proves macho enough to play too; and, finally,
5) the wages of defeat do not include the cost of an oil pan.

Since Pareto-optimality favours Proto-Chicken over Chicken-proper, it should not surprise us that selective pressure has renders most of us much more disposed to the former than the latter. [26] But, except for point (5), Proto-Chicken is really just a way of relocating the game of Chicken from its normal time and venue - i.e. after midnight a few miles from town - back a few minutes - i.e. before midnight - en route from the bar back in town.

10. How to Win at Chicken:

Now then, here is a way, prima facie at least, Column can virtually guarantee victory. There is an accessory on the market which, when activated from the dash, produces the following two effects. First, it activates a red strobe mounted on the roof of the vehicle - a strobe peculiar to just this accessory - and second, until the odometer clocks one mile, it locks both the accelerator to the floor and the steering wheel to the white line down the centre of the road. No doubt the manufacturers of these devices will have given them highly marketable names like "Victory or Death!", "Dare to be Boss!", or "Satan's Cruise Control!"

Now then, Column can reason, could it ever be irrational for me to activate my Cruise? Only if Row's vehicle is similarly equipped and Row's already activated his. But if I have already activated mine, Row would be a fool to activate his. So, just as in the ED - where the winner is whoever manages to draw first on her irrationality holster - so in the CD the winner is whoever manages to first activate his Cruise.

But, just as in the ED - where we saw that the pharmaceutical companies can always be counted on to have put on the market a pill to defeat the rationality of ingesting the last pill they have just put on the market - so in the CD we can rely, either on the same people who have been marketing the Cruise or, if not them, their competitors, likewise to have marketed a device designed to defeat the rationality of activating the Cruise. Consider, for example, the "Cruise-Buster" - it sells for a fraction of the cost of a Cruise! - consisting in a pair of eye-glasses which, when activated by a switch which also activates a blue strobe, allows the wearer to see all but red light. If Column's Cruise-Buster were flashing, could it be rational for Row to activate his Cruise? Obviously not. Unless, of course, Row had already activated his Cruise-Buster-_Buster_. A Cruise-Buster-Buster is a switch which either activates a Cruise-Buster-Buster-wearer's Cruise-Buster or does not; but the Cruise-Buster-Buster-wearer has no way of knowing which.

(The Cruise-Buster, by the way, is a self-blinder. The Cruise-Buster-Buster is a self-self-blinder.)

The moral of the story should by now be obvious. One need not imagine an infinite number of _kinds_ of devices to see that, so long as there is an infinite _number_ of devices, there is an infinite number of combinations and permutations of them. Nevertheless the _outcomes_ produced by these combinations and permutations can be reduced to only four categories, i.e. those wherein Column will emerge victorious, those in which Row will win, those wherein they will both chicken, and those in which they will both die. But _some_ of these combinations and permutations are such that neither players can _know_ which of these four outcomes will be produced. Is it irrational to opt for a strategy which produces such an uncertain outcome? Not if both players are prepared to risk the consequences, multiplied by the 25% chance each, of their mutually staying the course, mutually chickening out, or unilaterally chickening out, in exchange for the consequences, multiplied by the 25% chance, of unilaterally staying the course. So, it would seem, the CD is not deconstructable.

But if each of Column and Row are prepared to risk mutual face, unilateral face, and even _death_ in a CD in the hope they might thereby _gain_ face, _a fortiori_ they would be prepared in a _P_D to risk an extra one or three years in the hope of thereby gaining two. That is, if the PD is equivalent to the CD, then it follows that neither is the PD deconstructable. All that remains to be shown, then, is that, _in all respects germaine to the issue at hand_, the PD is equivalent to the CD.

11. _Prisoners' Dilemma is a Game of Chicken_:

> You and I, the 'prisoners', are separated. Each is offered the choice: to rat or not to rat. Ratting is done as follows: One reaches out and takes a transparent box, which is seen to contain

a device which will force one's co-player to chicken out in a game of Chicken.

> A prisoner who rats gets to keep the [device]. If either player declines to rat, he is not at all rewarded; but his partner is presented with

a device that will prevent his own chickening out in a game of Chicken,

> nicely packed in an opaque box. Each faces a long sentence and a short sentence to be served consecutively; escape from the long sentence costs

staying the course in a game of Chicken, "and escape from the short sentence costs" one's co-player chickening out in a game of Chicken. But, one need not be reminded,

> it is irrelevant how [we] propose to spend [our devices]. So the payoff matrix looks like this. If we both rat, we each get

a device which will force one's co-player to chicken out in a game of Chicken, with which, presumably, we will each buy off our short (i.e. two-year) sentence, leaving ourselves each with three. If neither of us rats we each get a device that will prevent his own chickening out in a game of Chicken, with which, presumably, we will each buy off our long (i.e. three-year) sentence, leaving us each with two. But if one of us rats and the other does not, then the ratter gets both the device that forces his co-player to chicken out in a game of Chicken and the device that will prevent his own chickening out in a game of Chicken - and so can afford to walk - while his partner gets neither - and so languishes the full five. "There we have it," I conclude, "a perfectly typical case of Prisoners' Dilemma."

Now then, I continue, "[m]y decision problem, in a nut-shell, is as follows; [and] yours is exactly similar.

1) I am offered [a device that will force you to chicken out in a game of Chicken] - take it or leave it.
2) Perhaps also I will be given a [device that will prevent my own chickening out in a game of Chicken]; but whether I will or not is causally independent of what I do now. Nothing I can do now will have any effect on whether or not I get [the device that will prevent my own chickening out in a game of Chicken].
3) I will get my [device that will prevent my own chickening out in a game of Chicken] if and only if

> you do not take your [device that will force me to chicken out in a game of Chicken].

The Chicken Dilemma, I claim,

> is the same as regards points (1), (2) and (3). The only difference - if such it be - is that

whereas in the PD I would rather serve no years than two, I would take two over three, and three over five, in a CD, by contrast, for some reason I would sooner serve five or three than two. But, as already noted, "it is irrelevant how [we] propose to spend [our devices]" So,

> inessential trappings aside [then], Prisoners' Dilemma is a verson of [the Chicken Dilemma], *quod erat demonstrandum*.

12. Asymmetry and Mixed Strategies:

But before I allow myself to draw what conclusions I may, I want to say a few words about the relationship between my self-cloaking devices and the role played in thinking about these issues by seemingly similar 'mixed-strategies'.

The essense of the symmetrist argument, recall, is that

1) each [player] knows each knows each is rational, as well as the information represented in the matrix; from which it follows that
2) each knows the other will be doing on his side of the wall/windshield whatever it is that the first is doing on his; from which it follows that
3) nothing can happen on his side of the wall/windshield that will not likewise happen on his; from which it follows that
4) only mutual defection or mutual cooperation are possible; from which it follows that
5) if one defects so will the other, just as if one cooperates so will the other; from it follows that
6) one can assure himself of mutual defection or mutual cooperation by, respectively, simply defecting or cooperating himself.

But, it has long since been observed, one can break this symmetry by precommitting oneself to abide by the outcome of say, flipping a coin, tossing a die, or drawing a card. Hence one can block the inference from (2) to (3) just in case it can be shown it can be rational to so precommit oneself. So, mutatis mutandis, the question remains whether it can be rational to adopt one of the similarly outcome-unpredictable strategies that I have been suggesting.

My answer was it could be, and would be, just in case Column and Row are prepared to risk mutual face, unilateral face, and even death in a CD, in the hope they might thereby gain face, or in a PD to risk an extra one or three years in the hope of thereby gaining two. For simplicity I gave each of the four possible outcomes an equal probability, i.e. 25%, which is mathematically and logically equivalent to tossing a coin. But, of course, die-tosses and card-drawings - and various concatenations thereof - offer more fine-grained options. For example, in a CD one might opt for a mixed precommitment strategy that produces a 1% chance of being killed and a 30% chance of being defeated, in exchange for a 2% chance of emerging victorious and a 67% chance of being embarrassed.

Of course how one distributes these probabilities will be a function of how he values each of the outcomes. Minimaxers will elect to minimize the chances of death at the expense of reducing the likelihood of victory and increasing the likelihood of defeat. Disaster-avoiders will prefer to avoid defeat at the cost of increasing the likelihood of death. And so on. The sole constraint on this kind of thinking, given the premises of the symmetry argument, is that I can never hope, nor need I fear, that you have elected a strategy which, in combination with my own, will produce disparaties in our expected utility matrices. That is, the probability of my emerging victorious can never be other than the probability of *your* emerging victorious. And so on.

Of course I can *try* to throw you off your calculations by *pretending* to be more macho than I am, either by posturing in the bar, or by staying the course on the straightaway longer than I would otherwise be inclined to do. But given the premises of the symmetry argument such posturing would be to no avail. For *you* are assured that the discrepancy between my pretense and the *real* level of my machismo is precisely the discrepancy between yours and yours. In fact the premises of the symmetry argument render pretense utterly redundant. So mixed-strategic thinking breaks the symmetry, and therefore the inference from (2) to (3), but only where it is rational to opt for such a strategy. And this, in turn, will depend on the nature of the expected utility matrices involved.

Self-cloaking, then, is little more than a thinly veiled version of mixed-strategic thinking. True, given the premises of the symmetry argument, by self-cloaking I simultaneously ensure you will self-cloak. But then surely I can estimate the probability of the various outcomes if I self-cloak before I self-cloak. So I can only self-cloak where it is rational to self-cloak. And, once again, this will depend on the nature of the expected utility matrices involved.

What follows, then, is only that some PDs are not NPs; which is precisely the conclusion reached by a similar line of reasoning by Howard Sobel. [27]

13. Independence Revisited:

We have already seen that the equivalence thesis fails and that therefore - save for those rare occasions where it is irrational to self-cloak - the deconstructability we conceded to the NP does not penetrate to the PD. Now we have just seen that - save for those rare occasions where it is irrational to self-cloak - neither is the PD deconstructable in its own right. We can conclude, therefore, that we cannot - save on rare occasions - think our way out of the PD. Thus we have to do something to get out of it. Instituting an external sanction to alter our preference-orderings cannot overcome the compliance problem. So some kind of internalism, more or less along the lines suggested by Gauthier, remains - as we suspected all along - the only game in town.

In the process of establishing this, however, what exactly have we learned about the independence condition? One thing we have learned is that it is inadequately expressed as mere causal independence. Causal independence is a necessary component of the independence condition but not a sufficient one. Supplementing causal independence must be epistemic independence. But what exactly does epistemic independence amount to?

We can answer this question in turn by answering yet another. And that is, How does Gauthier think we can extricate ourselves from the PD? Not unlike Hobbes he thinks we can do so by taking measures to allay our grounds for fear we might be our own wary opponent's unwary opponent. And we do that, say both Hobbes and Gauthier, by removing our grounds for hope that she might be our unwary opponent. And this can only be done by simultaneously insisting upon, and submitting to, mutual (transparency or at least) translucency. But would it do us any good if all this transparency revealed about the other was that she had adopted a strategy that may or may not result in her cooperating? Obviously not. So simultaneously insisting upon and submitting to mutual transparency must simultaneously preclude such a revelation. As indeed it does. And yet Leslie insists it need not preclude such a revelation, for the simple reason that no such revelation is possible. Moreover, he argues, assuming we are both rational, not only is no such revelation possible, neither is it possible that what will be revealed is that my co-player intends to defect. So simultaneously insisting upon and submitting to mutual transparency is utterly unmotivated because entirely redundant, since we are already transparent to one another.

What I have taken pains to show, however, is that such a revelation is possible. Indeed it is this very possibility that the simultaneous insistence upon and submission to mutual transparency is designed to thwart. So what the epistemic independence condition amounts to just is the possibility that one's co-player has opted for one of these aforementioned perfectly rational, outcome-unpredictable strategies. [28]

5. RATIONALITY

1. Why Need Our Schema Be So Broad?

It is one thing to ask a philosopher to put his presuppositional cards on the table. It is quite another to demand to know where he got those cards, and why he thinks he is entitled to them. In what follows, then, I want to make explicit some of what must be at least implicit in the game-theoretic reduction of our moral dispositions. And I want to show, against each of several lines of criticism of those presuppositions, that a plausible line of defense can be mounted. But I will not presume - nor do I think I can be required, here at least - to press these lines beyond establishing their plausibility.

To begin with, Gauthier's is a story about how interhuman morality might arise out of the pre-moral interhuman condition, or - lest the word 'pre-moral' smack too much of historicity, then - out of our *a*moral or *non*-moral 'original position'. As such he is only parenthetically concerned - as likewise only parenthetically concerned were Hobbes, Locke, Hume, and Kant - to make claims about

a) whether anything akin to morality could likewise arise out of the (pre-, or a-, or) non-moral relations between humans and animals, and/or between animals and other animals, or about

b) whether anything at least moral-esque could arise out of the (pre-, a-, or) non-moral interactivity between humans and machines or, for that matter, between one machine and another.

Elsewhere, however, I have been arguing that the logic of Morals by Agreement settles both these parenthetical questions in the affirmative. [1] That is, I have been claiming that if the Gauthierian story can ground interhuman morality, then it can likewise ground at least some human obligations to at least some nonhumans, both organic and inanimate, as well as some nonhuman - both organic and inanimate - obligations to humans and other nonhumans, both organic and inanimate. And so it is important, for me at least, to lay the groundwork for these attenuated applications of our findings. Much of this, I concede, may seem highly counter-intuitive. So, I must also concede, even if my argument for 'machine rights' goes through, all that follows is that we will have to either bite the bullet on some of these counter-intuitive obligations, or else cite the argument as a reductio against the Gauthierian schema.

In any event, what metaphysical presuppositions I will be attributing to Gauthier will be those, and only those, that are necessary to ground the logic of Morals by Agreement. Anything else he presupposes, no matter how explicitly he may presuppose it, will be simply ignored. Of course it might be that the metaphysical story Gauthier needs to tell to get his meta-ethical story off the ground is so innocent as to make our exploration of it almost banal. If so, so much the better for Gauthier's meta-ethical story. In any event, let us see.

2. Mentality and Preferences:

Recall from our schema that "morality is to be understood as ... a set of rational, strategic, intramental responses ...", from which it seems clear enough that whatever else a moral being must be it must at least be both 'rational' and 'minded'. One question one might be tempted to ask, then, is this:

1) Is the set of all rational beings a subset of the set of all minded beings? Or
2) is it the other way around?

But, of course, given that a moral being must be both, this is not a question we need ask. That said, we do have to ask

3) what it is to be rational, and
4) what it is to be minded.

So suppose our answer to (3) makes reference to minds, but our answer to (4) makes no reference to rationality. Then our answer to "Is it (1) or (2)?" would be "(1)". Suppose our answer to (4) makes reference to rationality, but our answer to (3) makes no reference to minds. Then our answer to "Is it (1) or (2)?" would be "(2)". Suppose our answers to (3) and (4) co-refer. Then our answer to "Is it (1) or (2)?" is that rationality and mindedness are co-extensive. Suppose neither of our answers to (3) and (4) co-refer. Then our answer to "Is it (1) or (2)?" is that rationality and mindedness are intersecting sets.

Why is this important? Because some critics of our schema concede that

1) human beings are minded, but deny that they are rational. Others allow that

2) certain machines, for example, might be rational, but scoff at the suggestion they might also be minded. And still others hold
3) both of (1) and (2).

But what all these critics have in common is the following reductio:

4) If to be a moral being one must be both rational and minded, then the set of all moral beings is empty. Since
5) there are moral beings, it follows that
6) at least one of rationality and mindedness must be unnecessary. Since
7) human beings are both paradigmatically moral and manifestly minded but only dubiously rational, and since
8) so-called rational machines are paradigmatically non-moral and only dubiously minded, it follows that
9) any reduction of morality to rationality, be it contractarian or Kantian, is wrong-headed.

So whereas all Gauthier is committed to is that human beings are rational, what I need to add - if, that is, I am to get a case for 'machine rights' off the ground - is that certain so-called rational machines are minded. What is needed, then, is a conception of mindedness which Gauthier could himself endorse, but which will nonetheless accommodate my own designs on such an attenuated application of moral categories. The following, I suggest, will nicely do both:

Let us say that by a 'minded' being we will mean

> any (not necessarily spatio-temporally contiguous) regularity (or postulate) with respect to which it behooves us, given our epistemic and computational constraints, to take up - what Daniel Dennett calls - the 'intentional stance'. [2]

There need be no suggestion here such beings have their own consciousness, their own qualia, their own subjectivity or, for that matter, their own intentional states, at least in the sense of 'having' that the realist (or non-reductionist) about intentionality might have in mind. But nor need there be any suggestion they do not. That is, there may well be some deep, further, other-mind-independent fact of the matter, a fact of the matter knowable, if not by the being itself, then at least by God. But the intentional stance, as such, is utterly indifferent to such facts.

Now the first thing to notice about such beings is that nothing is ever one simpliciter. In fact any proposition of the form, "X is minded!", or "X has a mind!" period, is not a well-formed formula. Strictly speaking well-formedness requires that "X [be] minded (or [have] a mind) to (or for) [some] Y!"

And the second thing to notice, as a corollary of the first, is that in saying that "X has a mind for Y!" we are saying as much, if not more, about Y than we are about X. If nothing else what we are saying is that Y's epistemic and computational constraints are such that it must take up the intentional stance with respect to X.

Note too that something can be intentional, in the realist sense of 'being', and yet not have a mind in our sense of 'having' one. For example, on the Sixth Day, when Adam was still alone with God, God had a mind for Adam; and yet Adam did not have a mind for God. This is because God's epistemic and computational constraints - She having none - are not such as to behoove Her to take up the intentional stance with respect to anything! But this is not to say Adam could not have been a minded being in the realist sense of 'hav[ing] been', whatever that sense might be.

But for contractarians morality is not just a feature of minded things but - as the notions of rationality and strategy imply - a feature of things which are preference bearers. But, we might reasonably ask, can something be minded - even in this instrumentalist reading of the term - notwithstanding that it has no preferences? Suppose, for example, we conceived of God as being completely indifferent to Her creation. What would motivate us to think of Her as having a mind? Nothing would nor could. So, it would seem, to have a mind is to have druthers. But is the reverse equally true? Can something have druthers without having a mind? *A fortiori* not. So, it would seem, to say a moral being is necessarily a preference-bearing minded being is simply redundant. But, we must still decide, are all preference-bearers rational?

Suppose I prefer Coke to Pepsi, Pepsi to Root Beer, but Root Beer to Coke. Then since, as we are about to see, a condition of my being rational is that my preferences be well-ordered, and a condition on my preferences being well-ordered is they be such that the preference-relation is transitive, aggregating and decomposing, I am obviously irrational. But on that account alone would we want to deny that I nonetheless *do* prefer Coke to Pepsi, Pepsi to Root Beer and Root Beer to

Coke? Of course if a condition of my being regarded as having any preferences in the first place is they be well-ordered, then, of course, not only am I irrational, I am also mindless. But to say I am mindless is just to say that your knowing I prefer Coke to Pepsi would be of no use to you in predicting, were only Coke and Pepsi available to me, whether I would choose the Coke or the Pepsi. And yet this is clearly false. So likewise is it false all preference-bearers are rational.

Now a preference, in turn, is a relation between desires. Once again, suppose God had desires, but these desires were in no wise ranked. Would anything motivate us to attribute intentionality to Her? Even to suppose God has only one desire - say, for the best of all possible worlds, and She has no desire at all for any world other than the best - is to suppose nonetheless She prefers the best of all possible worlds to all others between which She is indifferent. But were we to suppose there is only one possible world, then, I submit, it would be completely uninformative for us to attribute to Her a desire for that world. So, similarly, we might suppose I prefer Coke or Pepsi to Root Beer, but I am indifferent with respect to the Coke and the Pepsi themselves. But this would only mean that, were Root Beer unavailable to me, and had you reason to try to predict which of the two soft drinks still available I will choose, my preference for Coke or Pepsi over Root Beer would be uninformative. That is, there is a perfectly serviceable sense in which my behaviour could accurately be described as - mindless.

3. Freedom:

What then is rationality? As a first approximation, at least, let us say that by 'rationality' we mean:

> that feature of an organism (or system's) cognitive and intramental executive apparatus in virtue of which it tends to maximize on the satisfaction of its preferences.

Now let us look at each of the key terms in this sub-schema in turn.

Let us assume we are satisfied we have a fix on what is meant by an 'organism' (and a 'system'). According to our sub-schema, then, rationality is a feature only of organisms (or systems) of which it can be plausibly (i.e. instrumentally) said they both

a) entertain preferences, as already discussed, and

b) enjoy robust cognitive and executive capacities, both intramental and physical.

Of course we could confine ourselves to intramental executions in order to allow that someone could be physically incapacitated but not, on that account alone, a-rational. Suppose, for example, I have well-ordered preferences, I know what I must do in order to bring them to fruition, but, alas, I am either utterly paralysed, or else any reliable connection between my will and my body has been effectively severed. Am I not a rational being, and therefore not a moral one?

Regrettably I am not. I am not a moral being because I am incapable of moral agency. And I am incapable of moral agency because I am incapable of agency simpliciter. I may or may not be a moral patient. (As I say, I have argued elsewhere that I am not even that!) But this is not the subject of our current enquiry. In answer to our current question: a condition of its being rational is that the organism (or system) can at least tense the appropriate muscle, even if that muscle meets preference-frustrating resistence.

Of course this raises the further question of whether one could be rational, and therefore a moral agent, if she were in a straightjacket. But it strikes me that this is one of those debates - in this case in action theory and the philosophy of law - that William James so fervently and rightly disparaged. [3] As we will soon see, the intuition we want to capture is that

> for some X to be regarded as a moral being by some Y, there must be some Z with respect to whom it behooves X to take up - what we will eventually come to call - 'the moral point of view', as a result of which, X hopes, Z will cooperate rather than defect in her interactivity with X.

If, for some reason, X is herself incapable of either cooperating or defecting - because, _ex hypothesi_, she is incapable of doing _anything_, we will no doubt want to deny X is a moral being for Y. But whether X's failure to be a moral being for Y arises out of this incapacitation defeating her rationality condition, or whether we allow she is rational but the failure arises out of this incapacitation defeating some further condition, is a matter to be adjudicated with reference to which algorithm-description does the least damage to our linguistic intuitions in this and cognate domains of discourse. Thus the debate is entirely formal rather than in any wise substantive.

A case in point of such non-substantivity is the now-famous Hart-Fuller debate - since perpetuated by Ronald Dworkin [4] - about the 'correct' juncture at which moral considerations are to be injected into our inferences from law to obedience thereto. Fuller maintains that if it is immoral - and hence unworthy of obedience - then it is not a law. [5] Hart insists it can be a law notwithstanding it may be immoral, but if so we may be justified in disobeying it. [6] The substantive vacuousness of the debate becomes clear, however, when one sees that all that is at stake is just how keenly we want to hold on to the inference from, "It's the law!" to "So obey it!" Preserving or jettisoning such inferences may be no trifling matter. But algorithms are individuated not by the order in which their data is processed, but by their necessary outcomes. On debates such as these have countless rainforests been levelled and careers forged; but, alas, precious little philosophical distance covered.

4. Information:

Yet another issue that might have to be settled with reference to linguistic intuitions rather than any algorithmic import is just how reliable one's cognitive capacities must be to qualify her as rational and hence as a candidate for moral agency. And it was with the hope of settling this issue that I included the words "in virtue of which". For without this inclusion we would have to allow that even the most dull-witted organism (or system) could count as rational just in case it enjoyed an incredible string of good luck. Suppose, for example, I desire a Porsche, I do absolutely nothing to acquire one, but I am awarded one nonetheless in a lottery I have only inadvertently entered. Suppose my life is virtually an unbroken series of just such fortuitous coincidences. *Mutatis mutandis*, even the most epistemically conscientious organism (or system) could count as irrational just in case it suffered an incredible string of *bad* luck. Suppose I do absolutely everything by which I might reasonably expect to acquire my Porsche, but just as the salesperson is about to hand me the keys the radio announces the utterly unforeseeable news that we are at war with Germany and all German assets are hereby seized. Suppose my life is virtually an unbroken series of just such unfortunate coincidences.

Now it might be supposed the codicil "in virtue of which" solves this difficulty by insisting that my cognitive and intramental executive capacities be causally responsible for the satisfaction of my desires.

This might help; but it will not be enough. Suppose, for example, I desire health, wealth, love and success, and so I smoke, waste, alienate and procrastinate. But, as it happens, medical science turns out to be just wrong about the effects of smoking on the respiratory system, my self-induced poverty inspires the pity of a dying millionaire, my lover undergoes conversion to a religion that glorifies the suffering of abuse, and by my publishing aught my dean infers this must be because I am just too brilliant to bother. Note that in each case the desire is fulfilled not in spite of my stupidity but precisely because of it. So here the fitness relation between desires and actions definitive of rationality is satisfied, and it is satisfied by my cognitive and executive capacities. Yet it would be odd indeed to say I am therefore rational.

Now it might be hoped all that is needed to repair the position against such Gettier-esque counter-examples is a way of ruling out what action-theorists and philosophers of law call 'deviant causal chains'. So one solution is to require that my cognitive and intramental executive apparatus direct my actions in a manner which is reliably in the service of my desires. This too helps; but it still will not do. For suppose God, or some benign genius, has so arranged the world that my every belief is true and my every desire is satisfied; but it is my belief which causes itself to be true and my desire which causes itself to be satisfied, rather than the other way around respectively. One intuition tells us to bite the bullet and concede that if such were the case then I would be rational. The other is to insist the causal direction be reversed. But then suppose God, or some benign genius, has so arranged my every belief that it turns out to be true and my every desire that it turns out to be satisfied. One intuition tells us to bite the bullet and concede that were such the case then I would be rational. The other is to insist I would be rational only if I also believed this reliable connection between the world and my desires and beliefs about it obtained. But, one might ask, would this be enough, or need I believe as well that this belief was likewise reliably connected to the world? And if so, would this not lead to an infinite regress? So to block the regress, it would seem, one must settle, at at least some point up the regress, for this reliable connection obtaining whether I believe it or not.

Still, what we are left with is that two people, with identical beliefs, desires, and intramental executive capacities, can be rational and irrational respectively just in case the one is, as it happens, thus reliably connected to the world, while the other is not. And this may seem counter-intuitive. But the reason it seems counter-intuitive, I suspect, is that it seems unfair. But it can only seem unfair if we are allowing ourselves to load into our notion of rationality some kind of evaluation of the person. And this, I submit, is just to misconceive the role of the notion in our conceptual network.

5. Interactivity:

Here are some more implications of our view:

We have already noted that, for the Gauthierian at least, being a moral being arises out of interactivity. Now on whether there can or cannot be desires - in the realist sense of 'be[ing]' - and therefore preferences, in a being which is utterly solitary, I need not presume to pronounce. But certainly if morality arises out of interactivity, and if there is at least one moral being, then no moral being can be unaccompanied. This is not to say if there is at least one moral being then there must of necessity be at least two of them. It might be supposed Gauthier supposes as much. But - it will prove important for my 'machine rights' project to note - nothing in our schema requires the accompanying being be itself a moral being. That is, "No person is an island!" true enough. But this is just to say if there is at least one moral agent there must be at least one moral patient besides. Whatever the conditions on moral patiency might be, however, it remains to be seen whether or not they can be satisfied by something that does not simultaneously satisfy the conditions on moral agency. So, for now at least, all we can conclude is that the reader cannot be a solipsist and stay on board. However she can deny the existence of other minds, should she be so disposed.

That said, I doubt Gauthier need take on any more metaphysical baggage than this. I doubt, for example, he need pronounce himself a substance monist, a nominalist, or an 'ist' of any other metaphysical sort or on any other metaphysical issue. So while I trust this has not been so loose-woven a metaphysic that it has all been just banal, I trust too that it casts its net wide enough to catch all but the most recalcitrant ontological skeptic.

6. The Marxist/Feminist Objection - The Underdetermination Problem:

Having thus sketched the contractarian conception of human rationality, we are now in a position to entertain a selection of objections to it. The first of these, as already noted, is that contractarianism presupposes that our 'default' condition, so to speak, is essentially egoistic.

Now it is certainly true that for Hobbes it follows from our appetitiveness [7], the scarcity of resources to satisfy those appetites [8], coupled with the equality of our threat and vulnerability to one another [9], to be human is to be "competitive, diffident and vain-glorious" [10], and these terms - all three of which, note, refer to diadic relations - hold between all persons. But we need suppose nothing of the sort. For example:

Since Hobbes' time we have come to believe that when we came down from the trees, and thence into the commons, we did so with at least some of our fellow-feelings already in place. That is, the family unit - possibly even an extended family unit - was already extant. No doubt too the pecking order within these units came about by (what we might call) proto-Acquisition. But that is beside the point. The point is the proto-Sovereignty of the dominant male of the unit did not arise out of the conditions which Hobbes thought gave rise to Sovereignty proper, i.e. as a means of extricating ourselves from a PD. It arose, thought he, from "natural lust", by which he did not mean that 'copulation truces', so to speak, were negotiated between "competitive, diffident and vain-glorious" equals. [11] By "natural lust" what he had in mind was some kind of pre-contractual symbiosis. So, according to our schema at least, intra-familial sentiments are not moral ones. And this, understandably enough, strikes many feminists as odd, if not downright offensive. [12] For what it does, in essence, is take what many women take to be the very phenomenological paradigm of their moral dispositions - i.e. the care they take of, and feel for, their husbands and children - and relegate it to the status of the merely 'private', and hence non-moral.

Nor is this oddity and offensiveness repaired by more modern contractarians. Rawls asks us to imagine ourselves coming to the bargaining table knowing that once the veil of ignorance is lifted we will find ourselves more attached to some members of the community than others. [13] But while he supposes we might bargain away some of our rights to treat such members preferentially, e.g. inheritance - that, in other words, we might consent to exchange some of these special attachments - he does not suppose we might bargain away the very having of them. So, once again according to our schema, the having of such special attachments cannot itself be moral.

Gauthier, by contrast, has us come to the table knowing not only that we will have such special attachments but also with whom. In fact they are not the means but the beneficiaries of whatever success one might hope to achieve at the table. So, according to our schema, such attachments are not themselves moral.

But, counter many feminists, even if the protective net a man throws over his wife - a.k.a. his 'woman' - and his children is motivated solely by a concern to keep an object of lust close at hand and a labour force under his control, and even if a woman's concern for her husband - a.k.a. her 'man' - is ultimately motivated by a concern to protect herself from the lustful designs of other men and to ensure the feeding and protection of her children, surely her love for her children is not reducible to a concern to keep a labour force under her control. Surely mother-love is the very paradigm of a moral disposition. And yet, says our schema, not so! So, she has every right to ask, what - other than a blind commitment to a logocentrically pleasing schema, and a 'phallologocentric' one at that! - could possibly motivate such a devaluation of what is indisputably phenomenologically the very paradigm of a moral disposition to the ranks of mere appetite?! Surely the more reasonable route would be to tell a plausible story by which social morality is simply an extension of mother-love. [14]

Indeed. Unfortunately, only the most implausible such story can be told. For in the first place, the phenomenology of morals is not the phenomenology of familial love. Of course that alone would not defeat the 'ethics-of-care' schema, since the claim is only that social morality is derived from familial caring. More to the point, however, is that the ethics-of-care story, even were to be believed, cites no residual, extant, default perspective from which we can judge whether or not this extension of caring from the immediate family to the far-flung corners of the Earth was appropriate or just a foolish mistake.

A more promising suggestion, however, is that, contrary to Hobbes, the natural condition of human beings is moral, but that this natural condition is corrupted by historically contingent circumstances. The task of the meta-ethicist, then, is to counsel the human community as to the means by which that natural condition can be recovered. Psycho-analytic feminism, for example, holds that the conditions Hobbes mistakenly identifies as natural are in fact the product of male absenteeism from the process of parenting. Marxist feminists maintain that Hobbes' "competitiveness, diffidence and vain-glory" arise out of the exigencies of economic stratification. Postmodern feminists argue that this corruption arises out of an intentional or invisible-hand 'conspiracy' to manipulate the language by which we construct our world, and so our moral responses to it. Radical feminists opine that, morally speaking at least, men and women are simply different species, and that the solution, at least to female moral corruption, is for women to separate themselves from men.

Is there any truth in any of these claims? I suspect there is truth in all of them. But the question is not what dispositions might we have had in a state of nature, i.e. before their being acted upon by contingent psychological, economic, linguistic and procreational praxes. The question is whether or not these default dispositions, whatever they might be, are being properly characterized as moral ones. That is, it is entirely possible that, but for some of these aforementioned contingent social arrangements, we would have been very different from the way Hobbes suggests we were, and to some extent still are. And it might even be that but for some of these aforementioned contingent social arrangements we would be phenomenologically indistinguishable from the by-and-large-moral way we are now. Our claim, however, is that what it is to be moral is not to simply *be* a particular way, but rather to have *become* that way as a means of extricating ourselves from the consequences of being some other way.

Now this claim, I concede, will give us no end of troubles. First of all, how do we reconcile this 'as-a-means' condition with our insistence that contractarianism is a logical rather than an historical claim? "If," as I said,

> the world had come into being only yesterday, and human beings along with it with their (now standard) moral dispositions already in place,

would they cease to be moral dispositions in virtue of the fact there *was* no

> other way from which, as a means of extricating ourselves, we became this way?

Obviously not. So our claim must be that moral they are, and moral they remain, in virtue of the fact that

> if we happened *not* to already have them our schema could explain why, on purely non-moral grounds, we

> would be well-advised to acquire them with all possible dispatch.

But if that is the case, then likewise moral would be, and remain, dispositions acquired in ways completely independent of the 'as-a-means' condition. And yet this undercuts our claim to be able to distinguish between dispositions that are properly moral, and those which are merely features of the (pre-, a-, or) non-moral original position.

Let us see if we can resolve this underdetermination. Let us say that a disposition is a moral one just in case

> it arose - or in its absence it would have arisen - as a means of extricating the agent from the consequences of its absence.

Thus, for example, constrained maximization is a moral disposition, regardless of how it came about, because in the absence of a CM disposition - i.e. were one still an SMer - it would be rational for her to adopt the CM disposition. Mother-love, on the other hand, is not a moral disposition because, in the absence of mother-love it would not be rational to become a loving mother.

This may be small consolation to the feminist. She may still bridle at the suggestion that mother-love is pre-, a-, or non-moral. But, as we saw earlier with the dispute over the role of morality in law, she may be bridling not over any substantive issue, but rather over what she regards as a misappropriation of the term 'morality', a term from which she is accustomed to making certain linguistic inferences she is loathe to now have to abandon. But it can hardly count as a decisive defense against a charge of conceptual confusion that, "Dammit, here lie my linguistic intuitions about my own morality, and so here they'll remain!"

7. The Existentialist/Postmodernist Objection - The Privileged Perspective Problem:

What may prove a tad more awkward, however, is that as a corollary to this injunction against intuition-tapping we are driven to the boast that

1) in neither motivating nor approving our theory is any moral judgment of our own involved, or, failing that, to the concession that

2) no such privileged, amoral perspective on meta-ethical theorizing is possible, and/or (else) to the insistence that

3) none is needed.

And yet (1) is about as implausible as is (2) humbling and (3) unsettling. Let us call this the 'objectivity problem', a.k.a. the 'privileged perspective problem'. And let us see what, if anything, can be done about it.

The objectivity problem, as a challenge to our enterprise, comes in two strengths. Both deny (1) and assert (2); but the difference between them is that the 'strong' version denies (3) whereas the 'weak' allows it. But in allowing (3) the weak version is all the more challenging since it allows (3) only because it deems the very notion of a privileged, amoral perspective on meta-ethical theorizing to be unintelligible. So let us look at each in turn.

In denying (3) the strong challenge amounts to a species of skepticism, and so on purely transcendental grounds might simply be dismissed. This is precisely what skepticism deserves; but let us make these 'grounds' more explicit. First a little background:

I take it we take it that what we are doing here is philosophy. This is important, because there has been a great deal of litigation of late over whether we are violating someone's cherished copyright on the term. Those who tout themselves purists accuse inter-level reductionism of scientistic sycophancy. The same charge is often levelled against erstwhile philosophers-proper who have sold out to cognitive science. (I am speaking of course in voce.) Of course one can always take refuge behind the 'institutional' theory of what-is-philosophy. On this view, we inter-level reductionists - and cognitive scientists - are doing philosophy because, "Well, that's what it says on our departmental office door!" But I think this would be unconscionable cowardice in the face of even less conscionable fiatism. I think a more robust response is both called for and available.

What makes making sense of our human condition so devilishly difficult is that we must, at at least one node - through probably several - in our network of sense-making, pull ourselves up by our bootstraps. As Sartre quite unsingularly observed,

> If every metaphysics in fact presupposes a theory of knowledge, every theory of knowledge in turn presupposes a metaphysics. [15]

And what has been said of metaphysics and epistemology may likewise be said of metaphysics and values.

I have elected to presuppose a metaphysics. I have elected more particularly to try to understand human beings - including their going about valuing things - under the unabashedly metaphysical category of organisms (or systems). Accordingly, I take my 'scientism' - and that is certainly what it is - to be a substantive philosophical position, a position no less subject to philosophical accountability as any that oppose it. And, mutatis mutandis, I would say the same on behalf of those philosophers of mind who have putatively 'sold out' to cognitive science.

Now some people say that, notwithstanding the logical necessity of presuppositions or choices, to have presupposed or chosen some particular metaphysics - in my case the metaphysics of organisms (or systems) - is to have already presupposed or chosen the set of values that inform that choice. Just exactly how such choices are informed by our values is less important than that they are. So, for example, it has become increasingly fashionable of late to observe that those of us who do game-theoretic, inter-level reductions on ethics do so because we are male and/or white and/or members of the ruling class; and that had we been female and/or black and/or downtrodden, we would have adopted an altogether different schema.

I entirely agree. In fact it will be the substantive conclusion of the final chapter of this enquiry that:

> no less than our moral dispositions, our philosophical methodologies and presuppositions are themselves just - instrumentally rational, strategic responses to patterns of interactive, game-theoretic activity.

Unfortunately I could not have convinced myself of this unless I had already presupposed the explanatory power of this self-same, game-theoretic, inter-level reductionist schema. But nor, claim I, could its detractors!

Of course turning skepticism on itself like this is an old ploy. What it shows, and all it shows, is one can neither intelligibly articulate, nor consistently argue for, a skeptical position. It does not show the position is false. Skepticism is a modal claim. To show it is false one has to actually do what it claims is impossible. In epistemology this means demonstrating that something is known. But if the skeptic will not accede to the verification techniques employed in the demonstration, one can do little more than shrug and go her own way.

In our case this means showing that (1) is true. But how does one demonstrate the absence of a moral judgment? One cannot. So we insist on reversing the onus of proof.

But if the skeptic refuses to accept it, what can we do but shrug and go our own (misguided?) way?!

Some people think we can afford to acknowledge the con-tingency of our methodological situatedness, because we are thereby freed to flit and sample, digest and synthesize, and then pronounce with an authority otherwise denied us. This is certainly the verdict of common sense, and more particularly of those who have studied the history of ideas, or travelled extensively abroad. The synthesis of perspectives, or so we are told, conduces to objectivity. This will hardly satisfy the skeptic, of course, since, as she rightly points out, there is always a perspective that remains unsynthesized. Still, there is something comforting - is there not? - in seeing oneself as akin to an ant, setting out on its hyperbolic climb towards objectivity, knowing full well it will take several generations of progeny before even one degree of incline is covered.

It should come as no surprise, then, that what I have called the 'weak' challenge might have to be taken more seriously than the strong. For in asserting (2) but allowing (3), what it amounts to is a kind of 'contented skepticism'; content, that is, to deny us (1) and yet let us proceed with our story nonetheless - in much the way an adult indulges the what-ifs conveyed to her by a child. On this view, the recognition of the contingency of our methodology leads not to any wholesale skepticism about sense-making, but simply to an incommensurablism about the propositional content of our respective narratives. That is, all that follows from there being no objectivity, even in meta-theorizing, is we each have a story to tell, the success conditions for these stories are internal to them - so even on their own terms some stories are better than others - but in any event we cannot understand each other's stories.

In my view, however, this is a more pernicious doctrine than even skepticism. This would not be so were it not that we have to live inside each other's narratives. More particularly we have to live inside each other's political narratives. Dorff's political narrative, for example, was one in which Goldman could not survive, let alone flourish. The Myth of Zion, in turn, is one in which Palestinians fare not much better. So our interest in each other's stories is by no means idle or casual. Moreover, narratives have meta-narratives, i.e. stories about what stories can be told. Another name for this is - methodology. So methodological consensus is no less critical to human cohabitation than is the political consensus to which it gives rise.

In any event, I think the jury is a long way from in on what, if anything, will be the upshot of the hermeneutical turn in philosophy that is now so much in vogue. In the meantime the observation per se that we cannot remove ourselves from the language-of-understanding's great hermeneutical circle has about as much significance - for any philosophical enterprise other than contemplating the language-of-understanding's great hermeneutical circle - as does Hume's problem of induction for any philosophical exer-cise other than Hume's problem of induction. That can be as great or as little as one lets it. As for myself, I have trouble enough coming to terms with the radical contingency of my every breath; so I am unlikely to be any more upset by the radical contingency of my every thought!

We want (1). Let us try to get it. Again, first some background.

I take it we too take it that what it is to be human is, at least in part, for what-it-is-to-be-human to be an issue. Some philosophers make very heavy weather of this. More specifically they charge that in reducing human beings to a clatter of rational maximizers, game-theoreticians have already decided this very issue; and moreover have done so completely discounting this salient feature.

Of course one's being a rational maximizer need not be inconsistent with its being an issue for her what-it-is-to-be-human. Unless, of course, she has elected to resolve that issue in favour of not being a rational maximizer, and has the power to act upon that resolution. In any event, that such self-reflection and determination are peculiar to human beings I have little doubt. But just how 'salient' this is depends on what task one's enquiry is designed to serve.

My own interests, here at least, are elsewhere directed. Still, I would have thought the first question to occur to anyone interested in this feature - i.e. that what-it-is-to-be-human is an issue for humans - would have been, Why would an organism have such a feature? Some story would then have been told, perhaps about how it might have proven advantageous for it to have an 'asking' function; and since it was not disadvantageous for it to seek answers it did not need, considerations of economy dictated against imposing another function to set limits on this asking.

Of course this is just one theory - one theory about what it might be for what-it-is-to-be-human to be an issue. But, I concede, even if it were true, it would say nothing about the meaningfulness to the organism of the what-it-is-to-be-human issue itself. Nor for that matter would it say anything about whether the what-it-is-to-be-human issue is one the organism needs to resolve, or one about which there is simply "No harm in asking!" So the only charge we need incumbently answer is whether, in analysing human asking in (admittedly) much the way we would the mating rituals of lemmings, we have in any way begged or prejudiced this what-is-it-to-be-human question. And clearly we have not. Similarly, then, in presupposing a metaphysics which, in turn, allows us to characterize human beings as sophisticated but natural organisms, and thence to subsume the human activity of making-value-judgments as among the features of such organisms, we are not passing a value judgment on this making-value-judgments. Were we to do so we would already be doing what it is we are analysing rather than analysing what it is we are doing. So, I would argue, for the reason given above the onus of proof lies with those who would deny us (1). And, for the reason just given that onus remains undischarged.

I trust the reader is herself a member of a both-traditions philosophical community, and that she will therefore rightly surmise there is a history to both the substance of these remarks, and their tenor.

8. The Postmodernist/Deconstructionist Objection - The Prescriptivity Problem:

But we are not out of the woods yet. For even if it can be shown that our schema is as good as any and better than most - because, for example, what it sacrifices in linguistic charity, a.k.a. what Quine calls 'conservatism', it more than compensates for in what he calls 'modesty', 'simplicity' and 'comprehensiveness' [16] - what remains to be shown is that any such meta-ethical story could have any pre̲scriptive import. That is, even if we have shown there is methodological privilege in analysing our current moral dispositions from the perspective of some hypothetical original position, we have yet to show that this hypothetical perspective can in any wise sit in judgment upon those current moral dispositions. To show this we would have to show that in some robust sense we still are in that original position.

It will not do, for example, to simply toss out a tu quoque. True enough, the deconstructionist - of whatever ilk, be she marxist, feminist, hermeneut, or whatever - is in no better position to translate her insights into the genesis of our current desires into a program for reforming the conditions that sustain those desires. But this is just to say none of us is entitled to make prescriptive recommendations. The notion of a categorical imperative, we all agree, is simply oxymoronic. All prescriptivity takes a hypothetical format. So other than as grist for anthropological curiosity, of what ethical import is meta-ethics?!

But before essaying our own answer to this question, let us establish our tu quoque first. I will not pretend to have exhausted the possibilities; but I offer the following as a representative sample.

Suppose I became convinced that cosmologically prior to me there exists a God who wants me to prefer Pepsi to Coke rather than Coke to Pepsi. I certainly might bring myself to do so out of my fear of Her. But out of the mere recognition of Her cosmological priority?! Obviously not. Similarly, then, suppose ontologically prior to my preference for Coke over Pepsi, there is an X whose interests would be better served by my preferring Pepsi to Coke. Should I be any the more moved by this discovery?! Cosmological and ontological priority, it would seem, are ethically irrelevant. [17]

A (seemingly) alternative way to introduce prescriptivity is to observe that with each recursion of deconstructing our existing values we are freed from the historicity - or more generally the contingency - of those values in precisely the amount of that recursion.

That is, we imagine ourselves, if we can, as we were just prior to acquiring our existing values - or, more commonly, some particular value. We then discount the determinant we think of as having in some way violated our freedom in acquiring the value (or values) that we did. And then we choose again. Of course more often than not what we chose in the first place will turn out to have been overdetermined. That is, it is unlikely that my preference for Coke over Pepsi is informed by anything as nefarious as a class consciousness perverted in me by the ruling class, or by a deep-rooted resentment over having been prematurely weaned. But often enough we allow we would have chosen other than we did. And, often enough, that realization can be empowering.

The difficulties with this view, however, are at least threefold. First, which determinants we identify as having in some way violated our freedom are themselves the product of this reflective process. For that matter the process itself is informed by the presupposed value of freedom. So how do we bootstrap ourselves out of our own contingency without at the same time running the risk of playing right into the hands of the very determinants we hoped to escape?

Here is a not-so-trivial case in point. In a nearby community several women were recently arrested for baring their breasts in public to protest the conviction of another women for the same offense. Presumably patriarchy has an interest in reserving the female breast for private viewing. Unfortunately the protest was attended by several thousand men, all ogling.

"Does patriarchy want to reserve my breasts for private viewing, or does it only want me to think that's what it wants, so in defiance I'll bare my breasts publicly, which is really what it wanted all along? In which case, I should ... But then, maybe it wants me to think it wants me to think ..."

The logic of their predicament, and ours, it seems, is not just akin to, but identical with, Putnam's infamous brain in a vat - a brain very probably not in a vat but, having dared to suppose it nonetheless might be, languishes there for all eternity pitifully trying to think its way out again. [18]

Second, even supposing there is some way to bootstrap ourselves, there are any number of recursive functions that can be performed on the same inputs, any two of which could produce radically different results. For example, suppose a pedophile has identified his having been himself sexually molested as a child as the determinant which, in acquiring his sexual proclivities, violated his freedom. Here is one function:

> He waits to see - not just if but - how the deconstruction actually 'takes', i.e. whether he actually ceases to be a pedophile, and only then does he allow himself to perform the next recursion.

And here is another:

> He checks the 'book' on how people normally develop who have not been sexually molested as children, and if it tells him they almost never become pedophiles he embarks on his next recursion as if he were not.

Clearly these two functions could produce co-extensive results; but, just as clearly, they might not. So how might the function be further precisified? The mind boggles!

And third, the deconstructionist procedure asks us to do the psychologically impossible. It asks us to imagine ourselves not having the values we do. In some cases - i.e. where it is impossible to be indifferent - this also means imagining ourselves having values we do not have. And this makes about as much sense as asking someone to imagine herself being the victim of a Peeping Tom but without her knowing it. [19]

Among philosophers who take this problem seriously, some have decided the only solution is to return to some kind of Aristotelian naturalism, to some species of essentialism, or to some combination of the two. There are - there must be! - they claim, features about ourselves that are at once a) sufficiently shared to i) ground our having common cause and ii) preclude ethical and/or cultural relativism, and yet b) sufficiently peculiar to us to iii) individuate us and iv) ground the indexicality of our choos-ing. One such feature, for women at least, might be reproductivity. A woman can be other than she is in many respects, observes Alison Jaggar, but she cannot be other than a reproductive being. [20] For even if she is a failure as a reproductive being, it is as a reproductive being that she is a failure. In the ethics of reproduction, then, i) women have common cause. And given this commonality, ii) such ethics can be neither entirely personal nor entirely culture-relative. Furthermore, reproductivity likewise individuates a woman. It does so iii) via the non-fungibility of her children. [21] This is why iv) "Because they're my children!" counts as an explanation for why she might have saved her own children from drowning rather than those of another woman.

A similar strategy is employed by David Braybrooke in his attempt to resurrect the much-maligned distinction between mere druthers and bona fide human needs. [22] There may be all kinds of functions, he allows, without which we would still be something, no doubt; but we might not then be human beings. A human need, then - as distinct from a non-human, e.g. scorpion, need, or a specific-human, e.g. black-belt, need - is one in the absence of the satisfaction of which one cannot function as a human being. In satisfying those needs, then, i) we all share a common concern. And given this commonality, ii) the autonomies of cultures that ignore these needs need not necessarily be respected. But, it might be added, iii) one such human need might well be the need for individuation. [23] So iv) "I did it my way!" might be a perfectly good explanation for why I did not do it another way.

More generally, then, let us say that by 'naturalism' we will mean the view that:

> For each of our natural kinds what is entitled to perspectival privilege - i.e. to immunity from

> further deconstruction - is that in virtue of which we are members of that kind.

And by 'essentialism' we will mean the view that:

> Immune too from further deconstruction are those features in the absence of which we could not be individuated from other members of our kind.

Naturalism, then, is meant to take care of (i) and (ii), essentialism to handle (iii) and (iv).

The difficulties here are at least threefold. The first, needless to say, is that to try to get prescriptivity out of natures or essences is to court the naturalistic fallacy. I say "court" rather than "commit" because neither the naturalist nor the essentialist need be committed to the view that one ought to be true to one's nature or essence. On this more charitable reading of naturalism and essentialism, then, the intended corollary of Kipling's famous advise to his son, i.e.

1) "If you can do such and such and so and so ... then you'll be a man, my son!" [24], is not the imperative

2) "So do such and such and so and so!", but rather
   the complement for the bi-conditional, i.e.
3) "And if you can't, then you won't be one!"

Nor would it be fair to say that any such advice is therefore trivial. It certainly would not be trivial if his son wanted to become a man and was casting about for how to go about doing so. So though normativity may not be prescriptivity, it is not on that account any the less action-guiding.

But here is a second objection. The problem with natural kinds is not so much that there are none. Even so they are a convenient enough fiction, without which we probably could make no sense at all of the world. Rather the problem is that even if there were natural kinds, there would be too many of them. That is, the world can be carved up in any number of intelligible ways. So even if we suppose natural kinds are distinct from (innocently) 'gruesome' ones - like Levites, leap-years and Christmas trees - or socially manufactured gruesome ones - like nationality or race - if naturalism is to be of any help to us in adjudicating the defensibility of the ways we project ourselves into the future, we need to suppose too that we can recognize our own natural kinds when we see them. [25]

That said, it might be argued that only the most recalcitrant skeptic could deny that biological species are the very paradigm of natural kinds. So too are male and female. But before we privilege more fine-grained constituencies - e.g. Jewish, heterosexual, right-handed, tall - we first have to satisfy ourselves that neither this particular way of carving up the world nor our identifying our-selves with these cuts rather than any other, might not themselves stand in need of deconstruction. So, for example, can Jaggar be sure her picking out reproductivity as the core of female nature is the product of uncontaminated reflection and not of categories imposed on her by patriarchy? For that matter can she be sure her picking out her womanness as the core of her own nature is uncontaminated by patriarchal (or alternatively capitalistic) interests? It seems - does it not? - we are back to where we started from, i.e. desperately trying to bootstrap ourselves out of our infernal, Putnamian vat!

But let us suppose we are being unduly pessimistic; that not only are there natural kinds but there are also tests by which we can recognize them. By what test in turn, then, do we go on to identify our essences? Let us try this:

> If what we are asked to imagine ourselves as other than is preventing us from being able to so imagine ourselves, then chances are we have hit upon at least one of the features constitutive of the agent that can be privileged from further 'deconstructive surgery'.

Then the third difficulty will be this. As a Nazi Dorfi had no difficulty imagining himself masquerading as a Jew. And, in time, as a Nazi-masquerading-as-a-Jew he could imagine himself almost forgetting he was merely masquerading. From almost-forgetting he could imagine himself occasionally forgetting. From occasionally-forgetting he could imagine himself usually forgetting. And from usually-forgetting he could imagine himself forgetting completely. But from none of this does it follow that as a Nazi he could have imagined himself forgetting completely he was not a Jew. So either we are going to have to allow that

a) essences change - which sounds a tad oxymoronic - or else that
b) one can be found guilty of indefensible self projection notwithstanding she had no way of

knowing it - which sounds a tad unfair - or else that

c) one has incorrigible, or at least privileged, access to one's own essence - which sounds a tad mystical.

Moreover, any other test we might devise will suffer from at least one of these three difficulties.

This is not to say that neither naturalism nor essentialism might bear the seeds of an answer to our query. I think they do. I think our essences _do_ change; which is why I prefer not to call them 'essences'. And I think normativity, though not as psychologically satisfying as prescriptivity, is all the action-guidance we need. In any event, it will have to do.

But just before we quit taking applications and begin instead shortlisting them, let us allow in a few more possibilities.

The standard objection to the liberal conception of the self is that it is precisely that - too liberal. More often than not, however, this intuition arises out of concerns not so much for the moral defensibility of the self but rather the physical defense of others. This is why we answer, "Obviously not!" when asked whether the pedophile could take refuge behind the facticity of his sexual proclivities. But, the liberal reminds us, there are ways of protecting ourselves and our loved ones from selves that are less than socially acceptable other than making social acceptability a constraint on the defensibility of value-acquisition. Marxist and feminist communitarians fail to see this, says the liberal. And this is what makes the observation that "the self is socially constructed" - which by itself is merely trivial - so insidious when preached by them. (Once again I am speaking in voce.) That the self is socially constructed, argues the liberal, hardly countenances society appropriating to itself unstinted license to set about deconstructing and then reconstructing selves. That is to commit the naturalistic fallacy! [26]

Still, there could be merit in what liberals are saying without there being none in what is being said by communitarians. What the liberal is offering is an account of how selves and their values are to be allowed to play in a community committed to the respect of selves and their defensibly acquired values. What the communitarian is offering is an account of how those values are acquired in the first place. The communitarian allows that values defensibly acquired are to be protected with every resource the community can muster. And the liberal allows that values indefensibly acquired warrant no such community protection. So between the two, where is the beef? If there is any real beef between them it must be in their respective accounts of defensible value-acquisition. So let us look.

Liberals insist we keep ourselves absolutely value-neutral when judging the defensibility of the content of other people's values. They do, however, put great store in those values' pedigrees. If, for example, the value was acquired under false pretenses or duress, and if but for that false pretense or duress the value would not have been acquired, then its prima facie right to be respected is defeated.

Here is an example borrowed, and then adapted, from James Woodward. [27] Beatrice, who has been religiously indifferent all her life - and so medically quite 'modern' - has recently been kidnapped by the Loonies, brainwashed by them, and now believes a blood transfusion would endanger her immortal soul. And she so informs us, in no uncertain terms, as she is wheeled into the emergency room after an about-to-be-fatal(?) car accident.

Let us set aside for the moment concerns other than for Beatrice herself. What does respecting her defensibly acquired values amount to in this case? Saving her life or letting her die? No doubt the deconstructionists we talked about earlier would have us deprogram her first and then have her choose again. But, as we have already seen, this is just to beg the question since, let us suppose, she has likewise let it be known in no less uncertain terms she does not wish to be deprogrammed.

Some people would urge - Woodward calls this the 'welfare-based' account - we do whatever we think is in Beatrice's best interests. [28] The problem with this view, Woodward stands not alone in observing, is it renders people mere featureless repositories for whatever utiles we think constitute interest. [29] That is, the utilitarian could ignore her instructions just as readily if she had been a Loonie all her life. Others would prefer - Woodward calls this the 'consent-based' account - we do whatever Beatrice herself would instruct were she uninfluenced by the forces we think violated her autonomy. [30] The problem with this view, Woodward seems less quick to notice, is it renders people mere featureless repositories for whatever influences we think do not violate their autonomies. [31] That is, if Beatrice had been a Loonie most of her life, the Kantian could (rightly) surmise her autonomy must have been violated at some point in her youth. He could then give himself leave to violate her will just as surely as did the utilitarian.

But the problem with this view of these two views, Woodward acknowledges, is that without any further, independently principled constraints on our notions of 'interest' and 'violation', the two views collapse into one. [32] This is because anyone bent on saving Beatrice's life - "Because, dammit, it's in her interests!" - can simply pick out whatever it was in her history that made her think a blood transfusion would endanger her soul, brand it a violation of her autonomy, and then counterfactualize on that history until he finds - or, more accurately, invents - a Beatrice who would think as he does. [33]

To play the role of these independently principled constraints Woodward proposes 'depth' and 'longevity'. That is,

> a value is privileged from our counterfactualizing on its etiology, deconstructing it - call it what you will - in direct proportion to how "deeply held and longstanding" it may be. [34]

But how exactly do we plumb the depth of a value's embeddedness? Do we measure longevity in absolute time or as a proportion of time lived? How do we weight depth and longevity when they conflict? Woodward does not say. But these questions have a familiar ring to them, do they not? And what this shows, I think, is that the liberal account of the self is driven right back into the maw of the imponderables where we once had to abandon the deconstructionist, the naturalist, and the essentialist.

How then, by comparison, does the communitarian fare? Not much better, I fear.

We have just seen that the tension between our welfare-based (utilitarian) and consent-based (Kantian) intuitions was one we experienced within the liberal conception of the self, and more particularly within our strictly Beatrice-regarding concerns. The communitarian, by contrast, need experience no such tension, since she denies the Beatrice/non-Beatrice distinction that sets it up. That is, recall that the whole problem of what to do with Beatrice was premised by the claim that we can carve off from our body of concerns those which are purely Beatrice-regarding. This carving off is precisely wherein the communitarian sees our mistake. For by "selves [being] socially constructed" she does not mean selves are things that present themselves to the community seeking assembly, like an IKEA bookshelf come knocking at our door. Nor does she mean they are things constructed out of the material resources of the community and then learn to call themselves selves, like Pinoccio. By "selves are socially constructed" she does not mean selves are things at all. Rather what she means is selves are social constructs!

This is certainly an odd thing to say. But philosophy is no stranger to odd sayings. So, we must ask, to just how radical an ontology does "selves are a social construct" commit her?

First of all, that selves are distinguishable from the bodies with which we associate them was taught by both Plato and Descartes, twenty-four and four hundred years ago respectively, officially believed in the interim, and remains to this day very much at the core of our conceptual and linguistic network. Secondly, that selves may not be ontological primitives but merely composed of such primitives was taught by Locke three hundred years ago and has become, I would venture to say, virtually the 'received' view. So all that remains between the notion of a composite and a construct is, Who (or what) is doing the composing or constructing? If it is mind-independent reality then selves are things, and as such can be carved off from collections of things to which they might contingently belong. But communitarians are not the first to suggest it might be otherwise. David Hume thought so. So did Gilbert Ryle. [35]

That is, suppose that thinking of selves as things out of which community is constituted is as much a 'category mistake' as is thinking of desks, professors and students as things out of which a university is constituted. What makes a student a student, or a professor a professor, is not some or any intrinsic property of that student or professor, but rather a set of relations she bears with other students and professors. So might it be with selves. What makes me a self in the first place, and what makes me the particular self that I am, are my relations with other selves in the network of selves I call my community.

If so, what follows? One thing that follows is that the defensibility of a self's value-acquisition is a function of the defensibility of those relations with other selves that make the self a self in the first place, and make it the particular self it is. But then what the communitarian needs in place of a theory of defensible self-projection is a theory of defensible social-projection. That is, no less than the liberal - or, more generally, the individualist - with respect to selves and their values, the communitarian must, on pain of courting social nihilism, acknowledge that there are appropriate, more appropriate, and inappropriate ways by which societies might project themselves. And it is not at all clear the problems to be encountered by the communitarian in this regard are not precisely the ones we have already seen encountered by the individualist.

For example, suppose - indeed recall - we were once a society of slaves and slave-owners. We were, and to a large degree remain, a sexist society too. If, as a society of slaves and slave-owners, we asked how we came to be such, no doubt some story could have been told. Stories are told too about how our society came to be sexist. Some stories are more plausible and comprehensive than others. But what is prescriptively, or at least normatively, significant about these stories, other than that they are grist for deconstruction? But, as we have already seen, deconstruction is dangerously circular, functionally undecidable, and psychologically problematic, no less so for societies than for individuals. So, it would seem, with respect to the problem of prescriptivity at least, the communitarian is no better off than the deconstructionist and/or psychoanalyist, the naturalist and/or essentialist, or the liberal and/or individualist. So, one might wonder, why put on such ontological mileage for so little distance covered on the ground?!

Of course none of this is meant to settle the ontological issue, if there is one, between individualism and communitarianism. The point is simply that, regardless of how that issue is settled, we would still be facing this problem of prescriptivity. Nor, I suppose, need we settle the ontological issue. We could have a theory of defensible projection for individuals and another for communities. And though the two might conflict we need have no fear they could contradict, since the one would be addressing itself to the individual agent, and the other to the agency of the State - assuming, of course, there is such a thing. That is, there is conflict, but no contradiction, in saying to the State, "Impose such and such a law!", and to the individual, "Defy it!"

Still, it might be argued, the ontological issue nonetheless must be settled before we go any further since, in constructing our theory of defensible projection, the intuitions we will be tapping depend very much on whether we are thinking of individuals or societies. For example, one might have an intuition that the Mohawk nation ought not to forfeit its sovereignty, or that the State of Israel ought not to secularize. But it would be a mistake to allow these communitarian intuitions to inform our judgments on the defensibility of an individual Mohawk quitting the reserve for the city, or an individual Israeli opening his shop on a Saturday. The problem, however, is just this. We have all kinds of intuitions regarding the defensibility of individual self-projection. We have far fewer regarding communities; and those we do have are much less stable. Moreover, most of the communitarian intuitions we have are probably unreflectively transferred from our intuitions regarding individuals. So even if the communitarian turn is destined to take, it will take several generations of thinking in this communitarian way before we will have a critical mass of intuitions upon which to fruitfully draw.

My proposal, then, is that here, at least, we bracket the communitarian turn, take the liberal, individualistic conception of the self as given, and confine ourself to a theory of individual defensible self-projection. In so doing, however, I make no claims about whether any of our findings can be transported *mutatis mutandis* to communities, nor whether any would stand if it turned out we had chosen the wrong ontology.

So, after a representative half dozen failed attempts at trying to privilege some particular ontological feature of our original position, where does this leave us? Back where we started from. As the beings we are, regardless of how we came to be such. But this is clearly unacceptable. Why? Because to confine the grist for my rationality to the desires I have now on the grounds that it is irrelevant whether, from the perspective of the person I once was, it was or was not a mistake to have acquired these desires, is, as we will see, by parity of reasoning to regard as likewise irrelevant what desires I can now only anticipate myself some day having. Thus prudence, if such a thing there be, will have to be analysed strictly in terms of future referring desires which must nonetheless be only currently entertained. And yet such a conception of rationality, I shall argue, is utterly inadequate to the pursuit of anything even approaching a life worth living.

'Worth living' from which perspective? That is the question I take to be the task of the remainder of his enquiry.

6. THE MORAL DIALECTIC

1. Dispositional Pluralism:

In Chapter Three, recall, I argued along with Peter Danielson that constrained maximization - at least as characterized by Gauthier - is both undermotivated and incoherent. It is undermotivated because it would have one cooperate unnecessarily with an unconditional cooperator. (Nor can Gauthier defend himself with recourse to our Principle of Sufficient Reason, since in _this_ case the disposition in question engenders lost opportunities for exploitation, and so _does_ "exhibit an excess involving an additional cost that therefore counsels against it.") And it is incoherent because it traps itself in an infinite regress of mutual conditionalities. Both of these difficulties, I allowed, could be overcome by making the transparent having of some property R, and being transparently hardwired to cooperate with a co-player likewise sporting property R, mutually necessary and sufficient.

But, as Holly Smith pointed out - and as we will confirm momentarily - it is easy enough to transparently acquire property R. It is more difficult to transparently hardwire the requisite linkage. But unfortunately - once again given our cognitive opacity - it is impossible to do so without paying a price in vulnerability to exploitation and/or lost opportunities for it. Furthermore, Danielson's insight, recall, is that there is no uniquely rational solu-tion to the compliance problem. What cooperative disposition it is rational to adopt is a function of the agent-type equilibrium featured in the population one hopes to enter. And, notes he, it is a function whose output in most case is a disjunction. We can now see too that, given our cognitive opacity, neither is there a perfect solution to the compliance problem, regardless of which of the disjuncts one elects to adopt. So that there are psychopaths on the human landscape - i.e. people capable of exploiting us notwithstanding their having property R - and that there are 'dupes' too - and so opportunities for exploitation that we will miss - need not be an embarrassment for our solution. Ought implies can, as much for the rationality reading of 'ought' as for the moral. So we can only be faulted for counselling a solution if an alternative could fare better. And, as it happens, none can.

For example we might be tempted to think as follows:

Yes, the psychopath flourishes, but only in one-shot PDs. He does _not_ flourish in iterated PDs, because the moment he defects he undercuts his co-players' presumption that he too is saddled by the desired linkage. So we can cut the psychopath out of the social 'take' by making R

> some current property P, plus the _historical_ property of, once having acquired property P, never having defected against another player manifesting property R.

That is, we might be tempted to think that

a) the caveat "once having acquired P" is designed to encourage the adoption of property R by not penalizing a player for not always having had it; and that

b) we make it a condition of our cooperation that our co-player has "never defected against another player manifesting property _R_", rather than just property P, because

   i) just as Danielson does not want us to be transitively complicitous in the rewarding of

UD, so we do not want to be transitively complicitious in the encouragement of property P. And

ii) nor, I should add, do we want to be ununderstanding by penalizing a player for having taken perfectly reasonable defensive measures against exploitation by himself defecting on a suspected Grinch.

But, as I say, none of this is to any avail. For even this P± version of R does not protect our CMer from exploitation in the one-shot PD. This is because in the one-shot PD a would-be Grinch will satisfy the '+' condition, albeit trivially. So in the one-shot PD the P+ CMer will fare no better than the CMer-simpliciter. Moreover, as Danielson has argued - and he by no means stands alone in this - the iterated PD is an inappropriate model for the logical genesis of morality. Why? Because in the iterated PD one simply maximizes over a longer haul; whereas 'the moral point of view', by definition, must countenance a willingness not to maximize at all. Or, as Danielson puts it,

> Since iterated games can be solved by straightforwardly rational agents, they are not morally significant problems. [1]

So, he would no doubt argue, neither is P+ CM a moral disposition at all! So since we are to confine ourselves to one-shot PDs, and since when playing only one-shot PDs Grinches can invade a population of Rudolfs, whereas Rudolfs cannot invade a population of Grinches, it follows that Grinching is "a strategy more successful than" Rudolfing.

That said, we seem to be driven to the following inference: If

1) by a 'moral' disposition is meant, among other things, one which cannot be affected by the past behaviour of any co-player, and since
2) "human [beings] are not [sufficiently] cognitively transparent" to make property R the mentioning of the other's decision procedure in one's own decision procedure, it follows that
3) we can never hope to have a moral disposition which is invulnerable to exploitation by psychopaths. And so
4) by Danielson's reckoning my disposition not to cooperate with a known psychopath is not a moral one. [2]

And yet surely this is if not absurd then, at least, much harsher than necessary. Can we not instead hold on to our dismissal of iterated PDs as being irrelevant to the logical genesis of morality, and yet reject the much stronger (1)? Can we not make R some property _reliably_ - hence, contra (1), _inductively_ - connected (albeit indirectly) to the player's action?

2. _The Transparency/Translucency Move_:

Obviously we can since, just as obviously, we do. Of course R is no _simple_ property, like having a red nose. Rather, as already noted, it is the _complex_ property of at least

> _seeming_ to wear one's cooperative dispositions on one's sleeve,

- i.e. one's face, one's body movements, one's voice, and so on. Since I manifest this property myself, and since players I have encountered in the past who have manifested this property have almost invariably cooperated with me, I induce that this is a _reliable_ property, and hence that _this_ player, notwithstanding I have never encountered her before - nor for that matter do I expect to encounter her again - will cooperate just as I will.

Note, for example, that we are much less inclined to cooperate with a co-player wearing one-way-opaque sunglasses. Indeed, the fear that shaded military strongmen and enforcers strike in us is, I submit, parasitic on just this counter-transparency, self-cloaking ploy. Similarly, inquisitors strip their victims of their clothing while remaining dressed themselves, not only to create an asymmetry in physical vulnerability but also to create an asymmetry in body-movement transparency.

I am myself extremely uncomfortable leaving a message on someone's answering machine. Why? Because whereas her prerecorded message is not reactive to me, I am nonetheless being called upon to be reactive to it. The asymmetry disempowers me. So, in self-defense, I shrink back.

Note too that here we have a quite plausible explanation for both racism and linguistic zenophobia. Conventional wisdom notwithstanding, I do not think we need to be taught to be racists. Rather we learn by contact with other races that there is no need to be. For example, I remember as a very small child being somewhat uncomfortable with Afro-Americans. I suspect it was because I could not reliably read their facial features. I since lived for some time in an Afro-American community, and this discomfort almost immediately fell away. But even now I find that the inner-city Afro-American young male has cultivated a 'walk' which, I suspect, I subconsciously read from my lexicon of body language as signalling aggression - or if not aggression exactly then at least something akin to it.

I remember too as a somewhat older child being similarly uncomfortable with aboriginal Canadians, not because I could not read their faces or body language but because the (to me, at least, incongruous) stoicism with which they reported events that would have horrified a non-aboriginal Canadian - "I had a sister once but she froze in the granary!" - left me bereft of any point of emotional contact. This too fell away once I moved into a mixed neighbourhood. But I still experience this same disconnectedness - and hence this same hesitance - when speaking with an inner-city hooker or runaway, or when I pick up a hitch-hiker who has done 'hard time'.

Notwithstanding I am fluent in only two languages, I can, I think, read the tone of what is being said in virtually any European tongue. Moreover I have no difficulty getting past the "accents [put] on the wrong syllable' by speakers of ESL, regardless of their country of origin. But I have no idea what is being expressed when certain Middle Eastern women trill their tongues, nor, when people are talking in Chinese, whether they are happy, sad, angry or whatever. And so, once again, I shrink back.

Wincing from pain is obviously hardwired not so much to signal vulnerability - why advertize to a foe? - but rather the need for assistance from a friend. But we are not entirely born transparent to each other. We encourage transparency in ourselves, our children, and each other. So in circumstances where transparency is inappropriate we develop the counter-skill of a poker face. Moreover human beings be-ing social beings, we are also specialized beings. Thus young boys, who are being prepped for a world that will call upon more poker-faced responses, are taught that "Real men don't cry!" Pugilists need to be stone-faced for the reason cited above. But soldiers have this same stone-facedness drilled into them not because the enemy is likely to get close enough to smell their fear, but because their comrades might. That is, fear is contagious and so undermines morale.

So, in short, there are any number of factors that determine the range, the depth, and the plasticity of transparency, not the least of which being that in human lands - as distinct from Danielson's H*Land, where hardware and available computation time are literally weightless and infinite respectively - information and plasticity carry heavy maintenance and response-time penalties. [3]

3. Combatting Pluralism:

I have been urging, along with Campbell - and will continue to urge below - that the 'morality' pill it would be rational to take is the one that

1) produces a mark on one's forehead that only it can produce,
2) hardwires her to cooperate with those and only those with just such a mark on their foreheads, and
3) hardwires her not to take an antidote after one's co-player moves first in a sequential PD.

In so urging, however, I do not wish to suggest this is the only pill it would be rational to take. I accede to Danielson's point that the pill it is rational to take is a function of the agent-types in the population one plans to enter, and that this function can - and more often than not does - produce a disjunction.

At the same time, however, it should be noted that

1) the disposition with which one enters a population alters the nature of the equilibrium itself, thereby expanding or contracting the options for the next entrant. Since
2) one has to share the population with that next entrant, and the next, and so on, and since
3) not every equilibrium is equally productive in terms of the social profits for oneself, it follows that
4) in entering the population one _does_ have an interest in doing so with that disposition the entry of which will, in turn, encourage future entrants to bring with them dispositions which will, in turn, encourage yet future entrants to bring with them dispositions which will ... and so on ... maintain or _ob_tain the equilibrium with the highest possible social profit for oneself.

In fact one might even take the prudential route of absorbing a slight _dis_advantage upon first entering the population, in the hope of thereby encouraging immigrations which, over the long haul, will produce a higher social benefit for oneself than otherwise.

So, having acknowledged Danielson's point, I want now to set it (at least slightly) to the side. That is, rather than asking,

> For some population, P, what disposition (or set of dispositions), D. would it be rational for some would-be entrant, E, to adopt?,

I should like to ask instead,

> For this P, is there some (or some set of) D* it would be most prudent for this E to adopt?

Once again, cf. Danielson, I do not suppose D* will be the same for any P, nor even that D* will be a singleton for any particular P. Nor do I suppose a P's D* will necessarily be a smaller set than its D. But I do suspect that at most only some members of a P's D will appear in its D*, and that at most only some members of its D* will appear in its D.

That said, I do not propose (here at least) to defend my preference for seeking out a P's D* over seeking out its D against the objection that, since prudence does not require morality, D dispositions are moral ones, but D* ones are not. As with the 'straightjacket' and Hart-Fuller controversies we visited in Chapter Five, this may ultimately devolve to a mere verbal dispute. [4]

Nor do I propose to rerun Danielson's program, replacing his "find D for all H*Land P's" command with a "find D* for all H*Land P's" command. For even were I competent to do so - which no doubt I am not - Danielson himself acknowledges, as we have seen, that H*Land tournaments model only the "logic of cooperation and constraint"; and even at that they all but miss two of the most important dimensions along which human landscapes vary, namely hardware maintenance and response-time. [5]

And last but not least among what I do not propose to provide here is anything more than a 'guestimate' of what D* would look like for the human world. For in the first place, populations vary widely across the human landscape. So any observations I make are likely to be highly culture specific. Secondly, I lack the resources to perform the requisite quantifications on the features definitive even of my own local population to make much more than banal observations. And third, I have enough confidence in the descriptive power of the theory Gauthier, Danielson and I are advancing that I am generally content to take the 'low technology' road of lifting D* off such texts of 'accumulated wisdom' as the Bible, Polonius' advice to Laertes, Kipling's "If", and paternal aphorisms the like of, "The difference between a sport and a piker is a nickel." [6]

But this much is certain - and this is the point I want to belabour in the sections to follow:

> A conditional cooperative disposition's inflexibility, or resistence to 'toggling', varies directly with the assurance it provides, hence directly with the confidence it inspires, and hence directly with the likelihood of being rewarded by a co-player with a similarly highly inflexible conditional cooperative disposition. But it varies _inversely_ with the likelihood of being exploited by a co-player with a conversely _low_ toggle-threshold.

Which is just to say that given the two polar extremes of CM-proper (CMP) and _pseudo_-CM (PCM), a CMP will fare well with a fellow CMP but poorly with a PCM, and a PCM will do well with a CMP but poorly with a fellow PCM. In other words, everyone wants to play with a CMP, but no one wants to play with a PCM, including another PCM. So, in other words still, everyone including another PCM has an interest in excluding PCMs from the playground. She, of course, has no interest in excluding her*self*. But since the only way to exclude a PCM is to insist on transparently high inflexibility, the only way for a PCM to exclude *other* PCM's just *is* to simultaneously exclude herself. Thus "the logic of cooperation and constraint" offers *pressure* in the direction of CMP.

True, this pressure *relieves* the pressure exerted by PCM on UC. UC, of course, will likely have disguised herself as just another CMP so as not to toggle the RC *variant* of CMP. (So let us call her 'ucpaCM', the 'pa' standing for 'passing as'.) So assuming that RC incurs maintenance and response-time costs not borne by the *CC* variant of CMP, ucpaCM's passing exerts pressure on RC to simplify to CC. Of course the *very* low maintenance and response-time costs borne by ucpaCM are counter-balanced by the very high *passing* costs she must bear. So there *might* not be excessive pressure on RC-turned-CC and CC-proper to simplify even further to ucpaCM. On the other hand, *if* maintenance and response time costs are greater for RC, for RC-turned-CC, and for CC-proper than are the passing costs for ucpaCM, ucpaCM will have an 'ill-gotten' subsidy from RC, from RC-turned-CC, and from CC-proper. So, once again, no one wants to play with an ucpaCM, including another ucpaCM. So, again once again, everyone including an ucpaCM herself has an interest in excluding ucpaCM's from the playground. She, of course, has no interest in excluding herself. But since the only way to exclude an ucpaCM is to insist on transparently high flexibility, the only way for an ucpaCM to exclude other ucpaCM's just is to simultaneously exclude herself. Thus, as above, "the logic of cooperation and constraint" offers pressure in the direction of CMP. And since, as Danielson admits, "without the presence of UC", passing or otherwise, "[the] RC [variant] fares no better than [the more cost-effective] CC [variant]" [7], likewise does "the logic of cooperation and constraint" offer pressure within CMP in the direction of CC.

Which is what, and all, Gauthier suspected all along. But perhaps now we can see why he suspected it!

4. Combatting Recidivism:

Once again, I have been urging that the 'morality' pill it is rational to ingest is the one that

a) produces a reliable sporting of one's cooperative dispositions on one's sleeve - that reliability having been induced from the historical behaviour of others who have so sported - that only it can produce,

b) hardwires her to cooperate with those and only those with just such a reliable sporting of their cooperative dispositions on their sleeves, and

c) hardwires her not to succumb to the temptation of taking an antidote just before moving second in a sequential game or just after having been 'read' in a simultaneous one.

But so far I have argued only for (a) and (b). What about (c)?

Of course even without (c) - let us call the pill that produces (a) and (b) but not (c): PD:P1 - we have at least the beginnings of 'the moral point of view'. But I say only "the beginnings of" because I should now like to raise a third objection to Gauthier's characterization of the CM disposition. And that is that, even as it now stands, the effects of PD:P1 remains altogether too unstable to constitute anything approaching the phenomenology of the 'the moral point of view'.

To see this we need only note that, while Column is hardwired to cooperate with a fellow CMer, she remains epiphenomenally frustrated by her own hardwiring. That is, much as she *must* cooperate she would still really rather not. Thus *Row* might reason, quite rightly, that were there a second pill on the market - call it PD:P-1 - that could reverse the effects of PD:P1, and had Column reason to believe she had already elicited Row's cooperation, it would be rationally mandated for Column that she take the second pill. That being the case it would be rationally mandated for *Row* not to take the first one. And, *mutatis mutandis*, likewise would rightly reason Column. And so, it would seem, both players are back to where they started from.

Thus, we might surmise, the pill it is rational to take in the first place is not the aforementioned PD:P1, but rather a *third* one that has the *following* two effects:

> First, it produces a mark on one's forehead, a mark that can only be produced by ingesting this particular pill. And second, it hardwires her to *prefer* cooperation over defection with those, but only those, with just such a mark on their foreheads.

Call this, if you will, PD:P2.

But, as it turns out, even this is not quite what we need. For even though Column now prefers hers and Row's mutual cooperation over even her own unilateral defection, she nonetheless continues to prefer the consequences of her own unilateral defection over those of mutual cooperation. Thus Row might reason, once again quite rightly, that were there yet a fourth pill on the market - call it PD:P-2 - that could reverse the effects of PD:P2, and had Column reason to believe she had already elicited Row's cooperation, it would be rationally mandated for Column that she take this fourth pill. That being the case it would be irrational for Row to take the third one. And, mutatis mutandis, likewise would rightly reason Column. And so once again both players are back to where they started from.

Thus it is that we can conclude that the pill it is rational for Column to take in the first place is neither PD:P1 nor PD:P2, but rather PD:P3, i.e. the one that:

> produces the mark that only it can produce, and hardwires her to prefer the consequences of mutual cooperation over those of even her own unilateral defection with those, but only those, with just such a mark on their foreheads.

For though _from_ the post-ingestion of PD:P1 it is rationally mandated to ingest its antidote the moment one has reason to believe her co-player will (or has already) compli(ed), and though from the post-ingestion of PD:P2 it is rationally mandated to ingest _its_ antidote the moment one has reason to believe her co-player will (or has already) compli(ed), from PD:P3, by contrast, it is anything but rationally mandated to ingest its antidote, i.e. some PD:P-3, even if one has reason to believe her co-player will (or has already) compli(ed). For since Column now prefers the consequences of mutual cooperation over those of even her own unilateral defection, nothing remains by which she might be motivated to ingest PD:P-3.

Moreover, I submit, the taking of PD:P3, as opposed to PD:P1 _or_ PD:P2, captures the phenomenology of 'the moral point of view' quite nicely!

5. _The Empirical Inadequacy of First-Order Moves Alone_:

Or does it?

What I have shown so far - and _all_ I have so far shown - is that the pill that it is rational for Column to take in the first place must:

> produce a mark it and only it can produce - or at least virtually so - and be - or at least be reliably known to be - strong enough to virtually preclude any grounds, and therefore any chance, of recidivism back to straightforward maximization.

And yet _such_ conditions could be met, it seems, by taking the pill - or, as Kubrick's _A Clockwork Orange_ had it, having it rammed down one's throat by Acquisition - that:

> produces a mark it and only it can produce, and merely induces a crippling nausea; a nausea induced not only at the very thought of defection but, as well, at the very thought of taking the antidote.

And yet few of us would recognize _this_ as 'the moral point of view'. Indeed, all that distinguishes this kind of internalism from straightforward Hobbesian externalism is that

a) the compliance problem is overcome by making the dis-position to cooperate mutually conditional - i.e. by undermining the independence condition - and

b) the sanctioning mechanism, a.k.a. the Sovereign, is now located in the gut, so to speak, rather than outside the body.

But are we to conclude from this that, for a self- or other-disciplining disposition to count as 'moral', it must of (analytic) necessity produce contendedness rather than epiphenomenal frustration? I think not. For sometimes, at least, we do seem to experience epiphenomenal frustration in the face of our own moral dispositions. For example, as I have already confessed, I tithe not because I want to but because I feel I ought to. And, as I have also confessed, more often than not I would much prefer I did not feel as I do. One explanation, of course, is I am simply ill-ordered. Another is I am well enough ordered but I simply err in reporting, even to myself, that I do not want to tithe. And yet a third is I err in reporting, even to myself, that I feel I ought to. [8]

But a fourth way of making sense of my seeming senselessness on this score is to suppose that the strength of the exigencies of the game-theoretic genesis of tithing, whatever they may be, are not such as to require a robust contendedness; and that moreover there are probably countervailing exigencies - e.g. the unreliability of an adequate income for self-subsistence - that have counselled a certain plasticity to the tithing response. So - or so I shall now tentatively venture - what distinguishes a disposition which is phenomenologically moral, as distinct from one that is merely functionally so, has less to do with either its strength or its contendedness than with its second-orderedness.

6. (N>1)-Order Moves - Their Depth:

A case in point, I would venture, is fellow-feeling. If Column prefers the consequences of mutual cooperation over those of her own unilateral defection, and yet at the same time remains indifferent to Row's welfare and/or preferences, she might - indeed she eventually would - notice that there is something ill-ordered about her own preferences. That is, she might wonder why she prefers the consequences of mutual cooperation to those of her own unilateral defection notwithstanding her non-altruism and/or non-tuism with respect to Row. And so she might be tempted to take PD:P-3 after all, if for no other reason than to bring her preferences back into good order. Of course she can as readily bring her preferences into good order by dropping her non-tuism as by ingesting PD:P-3.

So what she needs is some reason to opt for the former rather than the latter. And, as we have seen, that she has - in spades!

Nor is there any limit to 'n'. Suppose that in order to elicit Mary's cooperation Paul must learn to care deeply about her. But much as Paul cares deeply about Mary who - for independent game-theoretic reasons of her own not shared by Paul - has learned to care deeply about some Peter who, as it happens, Paul - for independent game-theoretic reasons of his own not shared by Mary - just as deeply loathes. But, it should be easy to see, there are also independent game-theoretic reasons to care deeply about only those one judges to have good reasons for caring about the people they care about. So Paul has a choice. Either he can

a) forfeit his relationship with Mary - "I love you, darling, but I couldn't possibly marry you, because I could never trust someone who'd given her life to Jesus!"

b) learn to care deeply about this Peter - "Maybe I should take another look at this Jesus thing!",

c) drop the transitivity requirement on well-orderedness - "Look, you go to your church and I'll go to mine!", or

d) drop the well-orderedness requirement *tout court* - "Oh well, you can pick your friends and maybe even your relatives; but there's no accounting for love!"

Of course *which* of these options Paul elects will be a function of myriad factors, only some of which I have already rehearsed. But, I submit, what people are *doing* - when they are thinking long and hard, as they clearly do, about what to do in such situations - is precisely this: they are retracing their decision procedures, they are double checking their data entries, they are seeking to settle their network of desires and beliefs into a 'reflective equilibrium' that both

a) preserves as best it can the privileging of both
  i) its 'protocol' entries and
  ii) its core axioms, while
b) making the least revisions to the territory in between. [9]

So if, more generally, we can say that what accounts for a disposition's being *phenomenologically* moral is the depth of its embeddedness - and therefore the relative phenomenological inaccessibility of its rationale - is it any wonder that the phenomenology of 'the moral point of view' does not seem to manifest this instrumentality on its sleeve?!

This is not to suggest the phenomenology of a moral disposition necessarily varies directly with its instrumental embeddedness. For example, for those of us who practise some degree of pedagogical 'tough love', our doing so may be instrumental to our feelings of '_in loco parentis_' responsibility, notwithstanding the latter may be more robustly phenomenologically moral than the former. So perhaps the phenomenological recognizability of a disposition's being moral varies along _two_ dimensions: its instrumental embeddedness and its requisite inflexibility. But, I confess, these are mere speculations on my part.

In any event, by an '(n>1)th-order' disposition I mean:

> any disposition adopted to ensure the development of - and/or, once developed, the maintenance of - any ((n>1)-1)th-ordered disposition itself ultimately adopted in response to the exigencies of otherwise dilemmatic game-theoretic interactivity.

7. Their Range:

So far I have spoken only about just how many moves deep a second-order disposition might go. I should now like to add a few words about just how wide-ranging they can be as well.

As we have already seen, they can operate on

1) on beliefs - from the "Jesus 'Myth'" to "Jesus Saves!",
2) on desires - from that "Jesus stays dead!" to "Jesus be praised!", and
3) on preferences - from "Coke over Pepsi" to "Pepsi over Coke", or from "to defect rather than cooperate" to "cooperate rather than defect".

But, as we have seen as well, they can likewise operate

4) on constraints on decision procedures - from "insisting on the transitivity of well-orderedness in the embedding of objects of trust" to "dropping that requirement" (as in Paul's (c) option above),

or, for that matter, from "insisting on at least one's _own_ well-orderedness" to "dropping even that" (as in his option (d)). And they can operate

5) on _capacities_ - from "the capacity to access one's co-player's decision procedure in the CD" to "the _in_capacity to do so" via the Cruise-Buster.

For that matter they can even operate on the capacity to make _further_ second-order moves - from "the capacity to take BD:Pn" to "the _in_capacity to do so" via having already taken BD:P(n+1), as in Chapter Three above. And they might even operate on the capacity to

> retrace one's decision procedures, double-check her data entries, and seek a network in reflective equilibrium that both most preserves its protocol entries and core axioms and makes the least revisions to the territory in between.

That is, one might have reasons for disconnecting one's own self-monitoring module if, for example, one judges either that by so doing she might be able to perform tasks she would otherwise be too fearful to attempt -

e.g. William James' "leaping the abyss" - or that by _not_ doing so she will seriously undermine the assurance one's co-player may require as a condition of her cooperation in an enterprise of importance to her.

For example, my wife, being a committed Kantian, becomes markedly less cooperative the moment I betray virtually _any_ instrumentality in my thinking about our relationship. Understandably so. That is, to be both fair and honest, _none_ of us likes to be regarded as just another input in our co-players' utility calculus. So it is only in our co-_philosophical_ moments that I point out to her that both her Kantianism, and her predilection to exclude all but fellow Kantians from the 'playground', are precisely the moral dispositions one would expect from someone in whom the instrumentally requisite moral dispositions are most highly developed. Of course I cannot claim - nor do I - that the very fact of her being a Kantian _proves_ her Kantianism has been instrumentally adopted. For in the first place, she could as readily be a Kantian because she had always been. (In Chapter Five, recall, we called this the underdetermination problem.) And in the second place, even if she _had_ merely 'become' a Kantian, of what possible relevance could that be? (This, recall, we called the prescriptivity problem.)

8. And Their Flexibility:

That said, one must be cautious about the capacities one forfeits. In the James case one was being chased by a tiger; and so by short-circuiting her own self-monitoring module she had literally nothing to lose. The case is the cognitive analog of the chemical endorphins and adrenelin that suppress pain during and/or in anticipation of a life-threatening trauma. And in the case of myself and my wife - where our long-term mutual dependency counsels against either of us availing ourselves of what few opportunities for exploitation there may be - I have little to lose and much to gain. But, I submit, Socrates died only proximally from an overdose of hemlock. Distally, I suspect, he died from an overdose of defection-suppressant, which was itself probably induced by an overdose of algorithm-monitoring-suppressant. (Can there be a more convoluted way for one philosopher to call another a fool?!)

In short, then, 'moral pharmacology' needs to be as precise a science as anaesthesiology. As important as the suitability of the mixture, its internal compatibility, and the order of the administration of its constituents, is the strength of the mixture - to the logic of which I now turn.

7. THE LOGIC OF SELF-EFFACEMENT

1. Changing One's Mind (The Story So Far):

Each of MBA, AM, and my own undertaking, can be viewed as protracted entries in an advice column for psychopaths. And the message all three convey is - Change your mind! Each of us offers slightly different advice about what *kind* of mind to change into. Gauthier urges conditional cooperation; Danielson recommends any of the disjuncts that will not spoil the flavour of the population-in-equilibrium (or PIE); and I have been pressing for whatever will most sweeten it. But we are all of a mind that, conventional wisdom notwithstanding, nice guys - though, cautions Danielson, not necessarily the nicest guys - finish first!

Moreover each of us has a slightly different story to back up his slightly different advice. By Danielson's lights, Gauthier gives not enough thought to the payoff effects of dispositionally diverse populations. "The move to motivational dualism," says he,

> is a major advance away from the utopian monoculture that plagues rational and moral theory. It allows Gauthier's theory to begin to address the problem of partial compliance. Gauthier demands that rational moral principles prove robust enough to resist amoral predators. However we should ask,

> why stop at motivational dualism? What about the other sorts of agent - in particular other sorts of morally constrained agents? Gauthier does not consider this complication although it creates new problems for his theory. [1]

By Gauthier's and my lights, in turn, Danielson, because of his 'virtual' methodology, cannot but pay inadequate attention to the real-world costs of hardware burden and response time. No doubt Gauthier and Danielson would find myriad lapses in my attention too.

But neither Gauthier nor Danielson nor I have any advice to give the psychopath about how, exactly, to change his mind. Of course Danielson has no 'how to' problem, since his clients are free-wheeling software entities that can alter algorithms with the nano-pulse it takes to swing a logic gate. Gauthier's and my clients, by contrast, are less cognitively efficient and inter-cognitively transparent; so we have to push a 'pharmaceutical' logic instead. But whereas we tout ourselves pharmacologists, we are not pharmacists. We prescribe but we do not dispense. We do, however, offer referrals.

In any event, here, in a nutshell, is our story so far:

We take as our explanatory primitive that human beings are, initially at least, straightforward preference maximizers in search of the cheapest and most reliable way to extricate themselves from the dilemmatic patterns of interactivity paradigmatic of their interpersonal condition. Available to them is an infinite number of strategies, but each of them involves some kind of precommitment. As I say, how they manage to execute these precommitments is their business, not ours. Except that, since 'ought implies can' as surely on the rational reading of 'ought' as on the moral, if they cannot precommit, then neither will they be judged irrational for failing to do so. Also, our advice to them has been that internal mechanisms are generally cheaper than external ones, and more often than not as or more reliable. So if they can access an internal mechanism, even if an external one is available, we suggest they do so.

In very short order, our story continues, they learn that though cloaked strategies can be highly useful in games like the CD, transparent strategies are generally more suitable for games like the PD. They learn too that strategies that leave themselves epiphenomenally frustrated are highly unstable. And because they are unstable, they provide relatively low assurance to their co-players. So the cheapest and most reliable way out of the PD, they decide, is to discipline their co-players into cooperating by, first,

1) disciplining themselves to cooperate, conditional of course upon evidence their co-players _will_ cooperate conditional upon _their_ doing so, evidence which is itself only indirectly dependent on whether their co-players will cooperate in fact; and then,
2) making whatever other adjustments to themselves that may be necessary to retain or _ob_tain enough contentedness and well-orderedness that their own self monitoring subprograms do not toggle them, even if this means they might have to actually
3) sabotage their own self-monitoring modules to do it.

Thus the stronger and more direct the aforementioned dependence, the higher the assurance. The higher the assurance the more prevalent the cooperation. And the more prevalent the cooperation the more profitable the interactivity. On the other hand, the stronger and more direct the dependence the higher the vulnerability, because the higher too the stake-threshold at which one might toggle. So the 'trick', so to speak, is to seek out co-players with matchingly high toggle thresholds, but not _so_ high that one's _own_ threshold is therefore so high as to court disaster or preclude windfall.

Moreover, contra Danielson I have argued that, since this "logic of cooperation and constraint" is the same for all of us, we can expect a certain uniformity to the toggleability of our co-players. And what little diversity there will be will arise not so much from any parity of enterability enjoyed by threshold-diverse would-be immigrants, but rather from their diverse life experiences. These life experiences, in turn, will vary along two axes: their payoff records in past interactions, and the size of the stakes in the games in which they currently find themselves. And, of course, players adjust their toggle thresholds as they go along. So, for example, one speeds in inverse proportion to the frequency (and cost) of her tickets, and/or in direct proportion to the force of the grounds of her hurry. Similarly, classical revolutionary theory counsels the extermination of the ancien regime. [2] But such 'cleansing' will only be possible (or at least affordable) if unexpected. And if its completion is uncertain, it would be downright imprudent, especially if the new regime has yet to ensure its own security against the foreign erstwhile allies of the fallen regime.

George Mavrodes gives a similar gloss on the evolution (and sometimes devolution) of the immunity provisions in jus in bello. [3] Murdering POW's is only possible once they have been taken prisoner. But they will not be taken - and so the battle will be bloodier on both sides - if they suspect they might then be murdered. So violating the convention can be prudent if and only if either the violation can be cloaked or the violator is convinced the situation will not be iterated. This is why fidelity to the convention is almost invariably a good idea. But conventions are themselves just (sometimes more, sometimes less) stable behavioural equilibria. And so, like Danielsonian PIEs, they can sometimes be poisoned by unilateral predatory behaviour, or sweetened by unilateral restraint.

But, as already noted, the defensive manoeuvres I make against my co-players' togglings are a function of not only past performances, but also of both what they might have to gain by toggling and what I might have to lose by their doing so. So, for example, I seldom lock my car when there is only a few coins in the glove box. But I might take this dissertation with me - Why? Of what use would it be to anyone?! - if I stop for milk on the way to the post office. Likewise determined, mutatis mutandis, are the measures others will take against my own toggling. So, for example, your purse and confidences, I can assure you - but then I would, wouldn't I?! - are safe enough with me. But would I toggle for an unmarked million tossed over my fence by a fleeing bank robber? Hmmm ... So, as the saying (joke?) goes, "We already know what kind of people we are. We're just haggling about price!"

2. The Effacement/Self-Effacement Distinction:

What I want to do now, and for the remainder of this enquiry, is make some very heavy weather about interactive games the winning strategies for which involve the setting of peculiarly high toggle thresholds. But before I even begin, we need to be able to distinguish between cases in which it might be rational for me to impose - or elicit the aid of others to impose - peculiarly high toggle thresholds on myself, and cases in which it might be rational for me to encourage peculiarly high toggle thresholds in my co-players.

For example, notwithstanding they were merely obeying his own instructions, it was not Ulysses but his crew who tied him to the mast. So similarly, we might allow, there is a sense in which, being incapable of Institution, the Hobbesian savage consents to have Sovereignty imposed on him by Acquisition; just as a would-be-but-weak-willed quitter might consent to have his friends and family stop him from smoking, or a weak-willed-but-would-be suicide might voluntarily subscribe to an irrevokable suicide pact, or a weak-willed-but-would-be saver or tither might sign an irrevokable source-deduction contract. That is, there are cases in which, being desirous but incapable of self-modification, we solicit the aid of others to modify us. But these are all cases in which, presumably, the modification-agent and the modification-patient share common cause. Even in the case of court-ordered driver re-education one might cite such common cause, since the miscreant is, after all, free to refuse the re-education and forfeit his license instead.

But in the case of Winston in Orwell's 1984 - or, in the case of court-ordered sexual 'therapy' - the heretic and the pedophile (respectively) are not free to refuse. And in fact the therapy may be not only against his will and his preferences, but - albiet arguably in the case of the pedophile - even against his interests. In fact the reader may recall I alluded to this distinction in Chapter Three when I observed that

> making a soldier out of a civilian involves, among other things, adjusting just [such] thresholds. But one question is: Is it rational for us to so

> adjust him? Another is: Is it rational for _him_ to allow himself to be so adjusted by us?

More generally, then, a PIE consisting of CMers and SMers, concerned to lower its vigilance costs and increase its aggregate cooperate product, almost always has an interest in turning its SMers into CMers. But its doing so is almost never in the interests of an individual SMers so turned.

Now we could capture these distinctions by making _self_-modification the terminological primitive, and we could then coin derivative terms like 'assisted self-modification' for cases like the would-be quitter, and then perhaps 'imposed self-modification' for cases like court-ordered sexual re-orientation. But, I find, there is little of interest to say about assistance and/or imposition _per se_. And, as we will soon see, the exigencies of formalization suggest we make the definition of self-modification parasitic on, or a 'special case' of, modification in general. Accordingly, I proceed as follows:

3. Self-Effacement/Replacement Defined:

Let us say that by the '_effacement_', or '_replacement_', by some ('effacing' or 'replacing') agent A, of some feature f of some ('effacement' or 'replacement') patient S's (thereby 'effaced' or 'replaced') psychology, in some set of circumstances C1 - be S's f a belief, a desire, a preference, a capacity, or what have you - and so by f's 'effacement' or 'replacement' _by_ some feature r of S's ('effacing' or 'replacing') psychology in C1 by A, we will mean:

> the setting by A (with or without the assistance of some B) of the threshold at which, under C2 (i.e. the circumstances _replacing_ C1 reasonably anticipated by A), S's r will toggle back to S's f at such a height that, under C2, S's f is rendered virtually unrecoverable by (or inaccessible to) S.

Moreover, since we want to honour (in the way we talk) our (ontological) commitment to the view that it is _agents_ who are to be held forensically accountable for the objects they elect to appropriate or divest themselves of (in the 'special case'), or the objects with which they elect to saddle others or deprive them of (in the general) - that, in other words, mental objects, no less than physical objects like guns, are themselves morally (and hence rationally) neutral - we impose on ourselves the following rules of grammar:

If the subject of an otherwise well-formed subject-predicate formula is an agent, then the predicates "effaces" or "replaces" (or any of their cognates) can be modified by the word "irrationally" (or any of its cognates). But if the subject of an otherwise well-formed sentence is a feature, then the modification of the predicates "effaces" or "replaces" (or any of their cognates) with the word "rationally" (or any of its cognates) will be deemed redundant; and their modification with the word "irrationally" (or any of its cognates), will be deemed oxymoronic. That is, since it is agents who - sometimes wisely, sometimes not - efface and replace features, we will reserve the expression, "S's r re-places S's f in C1 for A!", to signify our judgment that the replacement of S's f by S's r in C1 just is rational for A. If on the other hand we want to say this replacement is irrational for A, we will require ourselves to say that "A irrationally replaces S's f by S's r in C1!" rather than "S's r irrationally replaces S's f in C1 for A!"

So, for example, we can say, as we shortly will, things like, "Kirk may have over-effaced Leroy's desperation!" But it would be redundant to say that, "Leroy's indifference rationally replaced his desperation!", since for some feature to have "effaced/replaced" another just is for it to have been rational for the agent executing the effacement or replacement to have done so, regardless of whether he did or did not do so in fact. And so, mutatis mutandis, it would be nonsensical to say that, "Leroy's indifference irrationally replaced his desperation!"

Now, as I have said, just exactly how one goes about effacing and replacing psychological features, be they one's own or others', is not our concern. So that the modification of oneself, or for that matter of others, may sometimes require the assistance of still others, will be given no more attention, here at least, than that the modification of oneself or others may sometimes require 'acquisitional' assistance, sometimes 'pharmaceutical', and so on. That said, we are now in a position to attend to the distinction between self-modification and the modification of others. And this we can do by making self-effacement a 'special case' of effacement in general. That is, by 'self-effacement' we will mean simply:

> the setting by A (with or without the assistance of some B) of the threshold at which, under C2 (i.e. the circumstances replacing C1 reasonably anticipated by A), S's r will toggle back to S's f at such a height that, under C2, S's f is rendered virtually unrecoverable by (or inaccessible to) S, where A = S.

For the remainder of this enquiry, as its title suggests, we will be focussing almost exclusively on the logic of self-effacement. We will be revisiting the logic of the effacement of others only when we ponder, as we will, the question of whether antecedents and continuers of oneself can or cannot be considered 'others'. [4] But in the interim, unless otherwise stated, the distinctions we are now about to draw within self-effacement should be deemed applicable, _mutatis mutandis_, to within effacement simpliciter as well.

4. _Distinctions Within Self-Effacement_ - _Stong/Weak_:

In _Reasons and Persons_ Derek Parfit unsingularly observes that SM - save, adds Danielson, in rare PIEs with "scrutiny costs high" enough for them to gain entry [5] -

is self-effacing. [6] But, I now want to add, it is also 'strongly' self-effacing. A feature of one's psychology will be said to be 'strongly' self-effacing just in case:

> it replaces itself with a feature which, were the original feature to remain, the network would be rendered ill-ordered.

And a feature is said to be 'weakly' self-effacing just in case:

> it replaces itself with a feature which, were the original feature to remain, the network would remain well-ordered.

The ill/well distinction, in turn, might be expressed in terms of 'erotic compatibility' and 'incompatibility'. If, in order to get what I want I only have to not want it, then insofar as my getting it and not wanting it are erotically compatible and hence well-ordered, I must only weakly efface my wanting it. But if on the other hand in order to get what I want I have to want not to have it, then insofar as my getting it and wanting not to have it are erotically incompatible and hence ill-ordered, my wanting it must instead be strongly self-effaced. [7]

Rather than flog our already overworked examples, let us try this:

Leroy is lonely, desperately lonely. What he wants more than anything else is to 'meet' someone. But he has learned from long and bitter experience - this gained at the local watering hole affectionately and, let us suppose, accurately referred to by the locals as "The Only Game in Town" - that the only thing less attractive to members of the opposite sex than a 'bar cruiser' wearing his desperation on his sleeve is that same bar cruiser wearing his desperation not to wear his desperation on his sleeve on the (shuffle of) his pant cuffs.

Leroy's best friend, Kirk, by contrast, boasts a perfect track record down at the Only. He offers to cure Leroy. (Notice that if Leroy accedes, this is a case of assisted self-effacement rather than imposed effacement.) Two hours of aversion-therapy - more or less along the lines suggested by Stanley Kubrick's A Clockwork Orange - followed by a cold shower, reinforced by a screening of Play Misty For Me through the VCR and, promises Kirk, "You'll be set!"

Notice that Kirk's prescription *borders* on overkill. Had he over-ordered, i.e. *three* hours of aversion, and rented *Fatal Attraction* instead, Leroy might have replaced his desire to meet someone with a desire to do anything but. Then the satisfaction of his erstwhile desire to meet someone would have been incompatible with the satisfaction of his *current* desire to do anything but. But, as it happens, Kirk knows his dosages. And, sure enough, he is subsequently rewarded by being asked to be best man at Leroy's wedding. So, in a nutshell, Kirk and Leroy owe their success to their noticing that Leroy's desperation is only weakly self-effacing.

Few of us, I trust, have lead youths so well spent we cannot even *imagine* such bar scenes. But some may have more difficulty imagining one in which the 'winning strategy', so to speak, might call for 'stronger' measures. True, it would be pointless for Leroy to visit the Only on a Thursday night, when the tables with the best vantage are reserved for women whose particular penchant is for 'turning' gay men. For had he become gay to win their attention, he probably could not have been turned. But even if he thought he *could* be turned, whatever the constraints we might ultimately elect to put on the rationality of self-effacement, it is far from clear they will end up sanctioning so dire a change in oneself, even if only so temporarily. We will have to see.

I feel compelled to add, in passing, that we should not judge the _women_ who come to the Only only on Thursdays irrational on the sole grounds that their penchant is unsatisfiable. For even if it were a penchant instrumentally adopted to satisfy some _prior_ penchant, we would have to know just exactly what that prior penchant was. Besides, it might well have been an 'original' penchant. But we would be ill-advised to make it a blanket condition on original penchants that they be satisfiable. To suppose otherwise - to suppose, that is, that it is irrational "to dream the impossible dream" - is, by extension, to suppose "it is better to be a pig satisfied than Socrates dissatisfied". That penchants be satisfiable _may_ be a imperative of reason. But _our_ sole commitment, recall, is that our preferences be transitively ordered. Moreover, "Ought implies can!" does _not_ in turn imply "Want implies can (get)!" [8]

That said, there _are_ women - or so I am told - who are particularly responsive to the challenge of a man who has taken a vow of celibacy. So suppose Kirk sends Leroy to the local seminary. Since the satisfaction of the desire to 'meet' someone, in the sense Leroy intends, is incompatible with the satisfaction of his _ex hypothesi_ recently acquired desire to keep his religious vows, desiring to meet someone could be strongly effacing if, but of course only if, he anticipates that he can be enticed to break his vows, and his current desire that his anticipated post-conversion self meet someone is greater than, or at least equal to, his current desire that his anticipated post-conversion self prove faithful to his vows.

I should add as well - and not just in passing - that one cannot simply map the strong/weak distinction - which is intended solely for the exigencies of erotic logic - onto doxastic logic and the logic of capacities. And yet, I have been insisting, beliefs and capacities can be self-effaced as well. How, then, do I propose to overcome this problem?

First, let us clarify the problem. Why are there so few atheists in the foxhole and/or so many deathbed conversions? Because, as Pascal pointed out, if there were a God - and so She could and (let us suppose, would) help out were She called upon to do so, but, let us suppose, only by a believer - the deathbed and the foxhole are precisely the times at which the cost of believing in Her (a few fervent Hail Marys) pales in comparison with the benefits of believing in Her if She exists. And this can be so irrespective of how unlikely

Her existence may be. Fair enough. But much depends on one's particular suspicions about the nature of God's psychology. He might believe She requires he become a full-blown theist; or he may believe She would be content with his merely acknowledging She _might_ exist. That is, the belief that there is no God and the belief that there is are doxastically incompatible. But the belief that as a matter of fact there is no God and the belief that there nonetheless _might_ be are doxastically compatible. [9] But if we take _this_ to be the strong/weak distinction for doxastic logic, God's demand that one's atheism be weakly self-effaced will amount to Her not demanding that his atheism be self-effaced at all!

And matters become muddier still in the case of capacities. By the strong self-effacement of my capacity to self-monitor, for example, is meant, presumably, my being utterly _in_capacitated; whereas by its weak self-effacement might be meant ... what? Its being merely crippled or diminished? But in what sense is being crippled analogous to being indifferent? And in what sense is either analogous to agnosticism? Can we afford to have a distinction - in this case between strong self-effacement and weak - the means for the distinguishing of which will vary in so _ad hoc_ a manner, depending on the ontological kind of the psychological feature involved?

The way out of this difficulty, I suggest, is to remind ourselves that, under the reading of rationality that drives our enquiry (i.e. Gauthier's and mine), the rationality of beliefs and capacities are a function of their instrumentality towards the satisfaction of desires (and/or, parasitically, preferences). That is, for Gauthier and for me, no less than for Hume before us, "reason is the slave of the passions!" So what we mean by the well-orderedness or compatibility of beliefs, capacities, or what have you, is just their instrumental efficacy in the satisfaction of desires (and/or, parasitically, preferences). And what we mean by the well-orderedness or compatability of beliefs and/or preferences, in turn, is just their instrumental efficacy in the satisfaction of the 'original' desires and/or preferences which give rise to them. But as to the well-orderedness or compatibility of these original desires and/or preferences themselves, we require only that they be transitively ordered.

So, for example, the reason why I cannot rationally believe simultaneously both p and not-p is not because I cannot psychologically simultaneously believe both p and not-p. I have not - or at least I need not have - any opinions about what is and is not psychologically possible. Nor is it because both p and not-p cannot be simultaneously true of the world. I need have no opinion about that either. Rather it is because I have induced, rightly or wrongly, that the co-entertainment of any p and not-p is not conducive to any network of beliefs which is in turn conducive to the satisfaction of any desires!

Am I prepared to entertain exceptions? I am not sure. But if not, no doubt I will be as ill-prepared to survive the rigours of 1984 as was Winston! Similar, then, are the reasons why I am generally ill-advised to incapacitate myself. Normally I need my capacities to get what I want. Indeed, that is why - in both the non-teleological, evolutionary sense of 'why' and our own highly teleological pharmaceutical sense of it - I have those capacities! But, as we saw with the ED and DD, there are exceptions to even this rule of thumb, exceptions for which I am well-advised to be prepared!

Desires, on the other hand - or at least some desires - cannot be regarded as purely instrumental in this way. Thus the only way the well-orderedness of all my desires could be self-effaced would be if I entertained some temporally prior desire, or some current meta-desire, which dictated the ill-ordering of all my desires including itself. A trivial case in point would be if I desired that all my desires be ill-ordered. But, as we saw in Chapter Five, we need not consider such cases since, the satisfaction of any such desire removes me from the domain of discourse.

5. Symmetrical/Asymmetrical:

The 'symmetry'/'asymmetry' distinction is parasitic on the 'original'/'default' distinction. So let us make the host distinction first.

Recall that when a feature of one's psychology is self-effaced, whether strongly or weakly, it is replaced, whether strongly or weakly, by something. (In the special case of its being replaced by nothing at all, it is in fact trivially weakly replaced.) This is why, as we have already seen, any proposition of the form, "f is self-effacing!" is not, strictly speaking, a well-formed formula. Strictly speaking well-formedness requires that, "f is effaced (or replaced) by [some] r!", or that "r replaces [some] f!" (In fact strictly speaking well-formedness requires that "f is effaced (or replaced) by [some] r in [some] C!" But more about 'circumstances' in a moment.)

6. But First an Aside About Stability:

But notice that it seldom if ever happens that "r replaces f!" and "f replaces r!" I say "seldom" because this symmetry, recall, is precisely what the symmetrist supposes about intending to defect and intending to cooperate when encountering one's psychological double in a PD. And I say "if ever" because I argued, recall, that one can break this symmetry by adopting a precommitment strategy to self-blind or to self-self-blind. Or, to put my case another way, note that the reason the symmetrist thinks intending to defect self-effaces to intending to cooperate is that she thinks by doing so one can cause (or quasi-cause) her double to intend to cooperate. But, I noted, in the wake of such a self-effacement have not the circumstances just changed? Intending to cooperate now self-effaces to intending to defect. (All the more reason, recall, to defect!) But then she is back to where she started from. So, I argued, the only way to prevent this recidivism is to find some way to precommit. And the only way to precommit is to alter either the outcome matrix (externalism), or the preference matrix (internalism), or both. So intending to defect and intending to cooperate are not mutually replacing. Why not? Because if the intention to cooperate would self-efface back to intending to defect, intending to defect is not self-effacing in the first place!

But let us consider instead Goldman and G-D. (To escape the awkwardness of my own self-imposed rules of syntax, let "Goldman", unless otherwise indicated, stand for the property of being Goldman; and likewise, mutatis mutandis, for "G-D".) Suppose we supposed that Goldman (the man) supposed that Goldman self-effaces to G-D, and that G-D almost immediately self-effaces back to G-D. If Goldman (the man) could have anticipated this, then, as with the symmetrist case cited above, he would have seen he could not hope to escape himself this way, and that therefore Goldman does not self-efface to G-D. In fact, we might say, Goldman self-effaces to G-D just in case G-D does not self-efface to Goldman! But now suppose, albeit improbably, Goldman (the man) had anticipated that, when circumstances dictated - i.e. were he still alive and were the camps about to be liberated - G-D would self-efface to Goldman and, moreover, that G-D (the man) could and would self-efface back to Goldman. Even if Goldman (the man) thought G-D (the man) might not self-efface, he knew - even if G-D (the man) himself would not - that he would not be executed by the Allies. So the situation would have been no different from Leroy replacing his desperation with his commitment to celibacy; assuming, of course, he could trust that that commitment, in turn, would itself self-efface in response to the sexual arousal the sexual arousal it was designed to engender was itself designed to engender.

Shall we conclude, then, that cases of mutual self-effacement are possible only in virtue of anticipated changes in circumstances, changes which are themselves _independent_ of the self-effacement itself? No, because the parry and thrust of adolescent courtship is precisely a case of mutually self-effacing and replacing dispositions which are such in virtue of _dependent_ circumstances. John loves Mary, as would Mary John, if but only if she/he does not love him/her, and desperately yearns/would desperately yearn for his/her affections to be returned. What distinguishes this case from that of the symmetrist argument, then, is just that in the former, but not the latter, presumably there is thought to be some value in the replacing 'moments', however fleeting they might be.

Such cases are, of course, rare. As are the Mobius strips and Escher drawings they structurally resemble. What I find interesting about them, however, is that they occupy a kind of logical no-man's land, halfway between self-deception and self-defeat - a land shared by Cartesian skepticism, perhaps?

7. Origins:

CM replaces SM in the PD, but SM does not replace CM, neither in the PD nor, a fortiori, in the circumstances now replacing the PD. Similarly, mild indifference on the subject of women replaces Leroy's desire to meet someone down at the Only tonight, but that desire does not replace his mild indifference on the subject of women. So let us say that by some 'original' position, O, of some feature r's replacement of some feature f, we will mean those conditions, C1, under which

a) r would replace f but, under C2 (i.e. the conditions created by r's replacement of f)
b) f would not replace r.

So if, for whatever reason, either Leroy-the-Desperate or Leroy-the-Indifferent were curious as to whether he was his own original or, so to speak, a mere derivative thereof, he has a way of finding out. Since

a) he would be Leroy-the-Indifferent whether or not he had been Leroy-the-Desperate or Leroy-the-Indifferent from the outset, but

b) it is _not_ the case that he would be Leroy-the-Desperate whether or not he had been Leroy-the-Indifferent or Leroy-the-Desperate from the outset,

his _original_ position is Leroy-the-Desperate. Which, _nota bene_, is _not_ to say - at least not without _independent_ grounds for saying so - that as a matter of _historical_ fact he was ever Leroy-the-Desperate, any more than it ought "peradventure be thought there was ever such a time or condition of warre as" _Leviathan_ describes. (In fact his original position could be Leroy-the-Desperate even if his historical position _were_ Leroy-the-Indifferent.)

For what reason might Leroy be curious about his own 'originality', in this strange sense of the word? One might as intelligibly ask, For what reason might I ask myself, while half-way up Tuscarora Mountain, in the midst of a downpour, pieces of lung hanging from my gasping maw and yoo-yoo-ing like silly-putty over my handle-bars, "What on earth am I doing out here?" I ask because I want to know whether or not I should _continue_ doing what I am doing!

And why do I say that the two questions are equally intelligible? Because insofar as many (if not most) of our _ways_ of being (and doing) are as instrumental to us as are most of our whats and wheres of being (and doing), it seems as reasonable to be constantly double-checking the appropriateness of the former as it does - as indeed we do - the appropriateness of either of the latter!

When must I perform this self-monitoring in this admittedly peculiar way? When the phenomenology of why I embarked on this arduous cycle trip across Pennsylvania may not be - indeed it seldom is - found in the phenomenology of the trip itself.

And why must I perform it? Because the phenomenology of why I embarked on this trip may not be - indeed it seldom is - to be found in the phenomenology of the trip itself.

But then neither is the phenomenology of the trip to be found in the phenomenology of whatever inspired it. Hell, had it been found there I never would have set off on this damn trip in the first place! But that can hardly mean that I should never have set off on it in the first place.

So, similarly, neither are the phenomenologies of Leroy's desperation and indifference to be found in each other. Nor, in cases of self-effacement, should we expect them to be. Indeed, we should expect them not to be. Were it otherwise the self-effacement obviously did not take. So, where we are trying to double-check our way of being, and where self-effacement is discovered as a constituent of our 'logical history' - i.e. that, whether or not we have independent grounds for confirming or disconfirming our suspicions about its involvement in our *actual* histories, there nonetheless exists a discoverable original position distinguishable from the extant one - we need to be able to 'come home again'. We need, as it were, to be able to 'return to our roots' - to our origins.

This, then, is (at least the beginnings of) our answer to the triad of questions we referred to in Chapter Five as the problems of underdetermination, prescriptivity, and perspectival privilege. For in the first place:

Although, as we are about to see, there may not be a *unique* original position, the *set* of original positions is nonetheless decidable. And, moreover, *certain* positions are decidably not among them.

Secondly: those original positions are the positions from which our existing position *may* be justified and/or critiqued. (The force of this 'may' will be explored momentarily.)

And third: these positions are privileged because, being themselves bereft of instrumentality, they may just *be* who and what we are.

In the face of which one might - indeed many do - shrink back in self-loathing. Nor will this self-loathing - if such one feels - necessarily be diminished by the knowledge that this feeling may itself be the product of an original self-love that has itself been simply self-effaced. In fact there seem to be two primitive existential responses to this 'self-recovery'. These, not surprisingly, correspond to two religious responses; which, not surprisingly in turn, correspond to two philosophical responses to the kind of meta-ethical enterprise that people the likes of Hobbes, Gauthier, and Danielson are pursuing. The one, it seems, is to celebrate our own instrumentality. The other, or so I am told, is to yearn to transcend it.

8. The Force of the 'May':

But, it might be objected, is there not a third response? And is it not just to ask, rhetorically, "Why bother with origins?!" For example, each of us came to philosophy by her own peculiar route. And each of us has concocted some self-serving story about that route. But suppose the truth of the matter turns out to be something like this:

Back in high school I was none too impressive on the football field. Neither was my best friend. So he and I used to hang around the door of the school library pseudo-intellectualizing in order to impress the female 'frosh' as they came in or out. We have long since discovered our own foolishness, of course. But in the meantime we became so intrigued by the issues we were discussing that in time we forgot _why_ we were discussing them. A quarter of a century later, should I give up philosophy and try to hone my football skills instead? Obviously not. And would the answer be any different if, in order to make our pseudo-intellectual strutting more credible, we had had to actually self-efface our true adolescent yearnings? A _fortiori not_. So, one might rhetorically ask, wherein lies the normative relevance of 'origins'?!

But, I counter, here is a more telling example. A woman wants nothing more than to "Stand By [Her] Man!" Of course in the absence of any independent reason(s) to question the originality (a.k.a. authenticity) of her feelings, she neither will nor should question them. After all, life is far too short to waste wondering whether one's preference for chocolate over strawberry is authentic or imposed! But suppose she _has_ such independent reasons. These can be that

a) something within her bridles against these feelings,

b) something within her does not square with these feelings, or

c) someone (perhaps a women's studies instructor at university) plants the suggestion in her mind that these feelings might not be 'original' to her.

Now unless something in (c) triggers something which triggers something, and so on ... which triggers either (a) or (b), (c) will be, and should be, irrelevant to her. In fact that (c) so often is relevant to women is itself strong evidence for (a) and/or (b)! So, similarly, with any information about how my interest in philosophy got started. That is, the self-deprecatory story I have just told is probably true, but in my case irrelevant.

So too would the story about G-D's origins be irrelevant to him. He certainly could, and no doubt would, allow that his psychology would be precisely as it is were he originally Dorff or Goldman; and that had he been originally Dorff he would not now entertain the psychology of Goldman. So clearly he is originally Goldman. But since Goldman was self-effaced - since, that is, when G-D looks at his rag-clad self in the mirror he sees a man in pressed SS uniform, and has in every other wise self-sealed his delusion - he cannot, nor ought he to be, triggered to double-check his own logical etiology. So if anyone is to be held accountable for failing to double-check his own past or prospective self-modifications, unfortunately that someone will have to Goldman.

So, in short, when are our origins relevant? When and only when there remains something within us that either a) bridles or b) does not square. Of course that nothing will do either is, by definition, virtually guaranteed by self-effacement. Which is why the instrumental fit between modified and pre-modified selves can only be grist for pre-deconstruction, so to speak, to the pre-modified self. On the other hand, when something bridles or does not square we have virtual assurance that something has been self-effaced, albeit clearly not all that successfully. Then and only then does it make sense to seek that something out. Of course it may - indeed most often it will - turn out that the self-effacement was well warranted, and that the residual bridling and/or non-squaring betokens only that the exigencies of the self-effaced network and circumstances were such as not to require a completely coherent replacement network. This, recall, was what we decided might be the case with the exigencies of the tithing response. That is, as things stand my being virtually hardwired to tithe flies in the face of some residual reluctance. But this, we conjectured, was probably because in the back of my mind I am holding the capacity to toggle in reserve, in the eventuality that times get particularly tough.

9. Residual Problems:

But I digress. What remain are two sub-problems of underdetermination *within* our answer to the problem of perspectival privilege. The first, as already noted, is that there may not be - indeed no doubt there seldom is - a *unique* original position.

Note that that there may not be a unique *replacement* position is not a problem for our view. For example, recall that after 32 recursions from a homogenous PIE of CC, reports Danielson, *any* of CC, RC and UC can replace UD in the PD. [10] Here is a similar non-problem. Until recently at least, it could be rational to trade in one's heterosexuality for either homosexuality *or* bisexuality in order to evade the military draft. Why is there no problem when the underdetermination works in *this* direction? Because there need be nothing upsetting about the observation that the way I am is an *arbitrarily* chosen response to the way I once was, so long as that choice was efficacious.

But, until recently at least, it could be rational to trade in one's homosexuality *or* one's bisexuality for heterosexuality in order *to* be considered acceptable by the military. Of course in all likelihood the (probably now homophobic) heterosexual soldier could not care less about his 'sordid' but now forgotten sexual history. But suppose I am a Sixteenth Century Spanish Catholic who has just discovered he is circumcized. And, of course, that the rest of my friends are not. Since fidelity to one's 'baptismal' faith is as important to Christians as it is to Moslems and Jews, it might make a difference to me whether my parents - who were obviously forced by the Inquisition to convert shortly after I was born - were Moslems or Jews.

Nor is this example entirely concocted. I have an adopted 'cousin' who was raised ultra-Orthodox, and who was well into her teens before she realized she was much too dark even to be Sephardic, let alone Ashkenazi. She eventually discovered she was Cree. And, notwithstanding her upbringing, she learned Cree, married Cree, *became* Cree. But suppose instead she had only been able to narrow her origins down to, say, Cree or Hispanic. What then?!

Of course origins involve not just *who* we once were but also how we were once situated. For example, until recently at least, it could be rational to trade in one's heterosexuality for homosexuality either to evade the military draft _or_ to make less odious a long prison sentence _for_ evading the military draft. But were I gay, and had I reason to believe my current sexual orientation may be the product of some kind of self-effacement, in thinking about what, if anything, I might _do_ about my homosexuality it _might_ make a difference to me whether I was originally motivated by a joint commitment to pacifism and my own personal liberty, or whether instead, having been arrested in the departure lounge for a purely _a_political evasion of duty, I had simply decided to make the best of a bad situation.

So, to summarize. What we hoped to capture by this notion of an original position is the (sometimes fat, sometimes thin) disjunction of psychological, material and game-theoretic circumstances under which it either was (historically), would have been (counterfactually), or would be (prescriptively) rational for one to replace it with the extant or recommended one. If it would likewise be rational for her to replace this extant or recommended psychology with this original psychology, then this original psychology is _not_, in this technical sense, original. For an original position to have normative force the search for it must be triggered. Once the search has been triggered and the position identified, more often than not the rationality of the extant position will be reconfirmed. (This 'reconfirmation', by the way, is the force of Gauthier's claim that MBA will by and large approve most of our considered moral judgments. [11]) But often enough too the extant position will be found want-ing. (This then, in turn, is the force of Gauthier's claim that MBA is more than a rubber-stamp on our considered moral judgments. [12]) But if it should turn out almost the entirety of our extant position is found wanting, we shall not - contra Rawls (and his conservative reflective equilibrium methodology - take this as conclusive evidence that it is the theory itself, rather than the bulk of our considered moral judgments, that is wanting. [13] This is the respect to which feminists err by not seeing that - Gauthier's right-wing, and Rawls' left-wing, substantive conclusions about distributive justice notwithstanding - it is Gauthier, not Rawls, who is their methodological ally. For it is he, not Rawls, who can allow that, insofar as female consciousness is informed wholesale by patriarchy, women might very well want to rethink themselves in their entirety!

But, after all has been said and done, we are left to concede that an extant position may have more than one original. Of course if each of these is such that the extant position, as riddled with niggles and ill-fittings as it may be, would be justified nonetheless, then this underdetermination is irrelevant. But otherwise, I concede, we are saddled with what for now, at least, I take to be an ineliminable difficulty.

Now our second problem, in turn, is that honing the aforementioned residual underdetermination is made all the more difficult by the fact that the units of what has been self-effaced, and/or of what has replaced it, may not be single features, like Leroy's desperation and/or indifference about women, but entire complexes of features, like whole personalities. So how does one take an entire personality profile and employ this 'transcendental' procedure to reconstruct the entire personality that gave rise to it?

To be sure this is a practical problem, but not, I think, a precedural one. Philosophers of biology encounter the same difficulty when trying to fix upon (what Ruth Milliken calls) the N-normal function of our biological adaptations. The N-normal function of an evolutionary adaptation is that function in the absence of the performance of which the adaptation would not have survived in the gene pool. But then surely the N-normal function of lips is as much to keep the ends of our mouths from fraying as it is to grasp food (or at least teat) and to talk. So, similarly, in addition to his psychological stresses and privations, Goldman may have had an insufferable barracks-mate he would have liked to treat as Dorff was treating him. So perhaps it was only the _combination_ of the two motivations that engendered his self-effacement. Here our Principle of Sufficient Reason will do some work. But clearly not _all_ the work that needs to be done. But, just as with fixing on N-normal functions in the philosophy of biology, one does the best one can, which is all that one can do.

10. _Defaults_:

In any event, by a 'default' position for some feature of one's psychology in some extant, postulated or recommended set of circumstances, I shall mean: any of those conditions in which one would find herself were that feature simply removed.

Assume for the purposes of this example that "If you're not part of the solution you're part of the problem!" and "If you're not part of the problem you're part of the solution!" Then the default position for feminism is patriarchism. And the default position for patriarchism is feminism. But the default position for

Leroy's post-operative indifference is not necessarily his erstwhile desperation. This is because in the absence of indifference one could be desperate or averse. Moreover among the default positions for G-D is probably not Goldman. In the absence of being G-D, erstwhile G-D would no doubt flounder about for awhile, like any other irremediable amnesiac, and eventually settle on something more or less non-descript. So the sets of original and default positions, though they may be co-extensive or intersecting, they need be neither. In fact they could be mutually exclusive!

In any event the difference, in a nutshell, between seeking one's original position and her default one is just the difference between wondering "How might I have got here?" and "Where might I go if I weren't here?" And so, in a nutshell, the difference between asymmetrical gettings-hither and goings-thither and symmetrical gettings-hither and goings-thither, will be the difference between being able to, and not being able to, as it were, get back home. So let us say that if the sets of original and default positions are mutually exclusive, then the original position was 'asymmetrically' self-effaced. If the two sets intersect then the original position will have been 'weakly symmetrically' self-effaced. And if they are co-extensive then it will have been 'strongly symmetrically' self-effaced.

Now, as I say, what I want out of the original/default distinction is to be able to clearly distinguish between asking:

a) "What might my circumstances have been such that it would have been rational for me to have become this way?" and
b) "What would my circumstances be were I suddenly not this way?"

And so what I want out of the symmetry/asymmetry distinction is to be able to mark off cases in which

c) (at least part of) the answer to (a) above is, "They'd be such that it would be rational for me to once again become this way!" from cases in which
d) (c) is not the case.

Thus, for example, Gauthier, Danielson, and I have been arguing that, under the game-theoretic conditions we have postulated at least, the answer to the question,

1) "What disposition might we have had such that it would have been rational to replace it with our current CM disposition (of some variety)?"

the answer is - "SM!" But none of us are particularly sanguine that "SM!" is the answer to the question,

2) "What disposition might we have were we suddenly _robbed_ of the CM disposition (of whatever variety) we now have?"

So, we might want to say, under the game-theoretic conditions we have postulated at least, CM is (only) _asymmetrically_ replacing of SM.

ii. _Contextual/Categorical_:

As already noted, _strictly_ speaking well-formedness requires that "f is effaced (or replaced) by [some] r in [some] _C_!" But, we might want to know, are there any desires, preferences, beliefs, capacities, dispositions, or what have you, which are _categorically_ self-effacing, i.e. self-effacing in _every_ context? Certainly none come readily to mind. For example, I opined that

> I cannot rationally believe simultaneously both p and not-p because the co-entertainment of any p and not-p is not conducive to any network of beliefs which is in turn conducive to the satisfaction of any desires!

So, provided I had any desires, if I did co-entertain some p and not-p that belief would probably self-efface. But would it self-efface in every context or 'possible world', so to speak? To which question, recall, I announced myself agnostic. Still, let us keep the contextual/categorical distinction, if for no other reason than to allow those who think that

> the co-entertainment of some p and not-p is not necessarily counter-conducive to a network of beliefs which are in turn conducive to the satisfaction of any desire

to express that conviction by simply denying there are any beliefs, capacities, dispositions, or what you, that are categorically self-effacing.

As to whether there are any desires, preferences, beliefs, capacities, dispositions, or what have you, which are categorically _not_ self-effacing, i.e. self-effacing in _no_ context - I am not sure. Could there be a world in which it could be rational to strongly efface each and every one of my current desires - including any desire I might have that my continuer have any of _his_ desires satisfied - if it were a condition of my having _any_ of my current desires satisfied - including any desire I might have that my continuer have any of his desires satisfied - that I strongly efface each and every one of my current desires - including any desire I might have that my continuer have any of his desires satisfied? Offhand I would say not. (But I admit this is a hard one to think through.)

12. _Mediated_/_Direct_:

Last but not least: In the same way, and for the same reasons, as there may or may not be a distinction worth drawing between features that are categorically self-effacing and those that are only contextually so, neither _may_ there be a distinction worth drawing between features whose being self-effacing is mediated through some further feature and those that are self-effacing directly, or in their own right. For example:

SM is (more or less) directly self-effacing in the PD. Why? Because it is SM that drives the SMer's effacement of his own SM, and it is his SM which the SMer effaces. But now consider the case of Snodgrass the Snob. Snodgrass, not unlike Leroy, is keen to meet someone down at the Only; and, not unlike Kirk, he has long since self-effaced his desperation. But what stands between _him_ and success is that he so utterly loathes the bar scene he cannot bring himself to go. So here, it would seem, what stands in need of effacement is his snobbishness, whereas what is driving that effacement is his keenness to meet someone. So, we might say, Snodgrass's snobbishness is only 'mediately' self-effacing.

The difficulty with this distinction, of course, is just the difficulty with analysing any claim which reduces to a counterfactual. True, in the absence of the desire to meet someone he would have had no need to rid himself of his snobbishness. But, just as surely (albeit tritely), in the absence of his _snobbishness_ he would have had no need to rid himself of his snobbishness. So why not say his _snobbishness_ drives its own self-effacement? Because, presumably, the keenness and the snobbishness are severally necessary, but only jointly sufficient, conditions for the self-effacement of the snobbishness, whereas ... what? [14] But rather than digress to several chapters on the problem of counterfactuals, safer, I suspect, to simply move on.

13. Self-Effacement and Its Cognates - Self-Defeat:

Having drawn a number of distinctions within the notion of self-effacement, I want to conclude this chapter with a few words about what self-effacement is not. That is, I want to distinguish it from two of what are, I claim, its mere cognates. First, self-effacement is not self-defeat. And second, though it is akin to it, neither is it self-deception.

For in the first place, self-effacement and self-defeat are properties of two very distinct ontological kinds - the former psychological features, the latter propositions. A proposition is said to be self-defeating just in case for it to be true it would have to be false, the paradigm case being, "It can be known that nothing can be known!" But nothing of the sort can be said of psychological features, because capacities, dispositions, and so on, are not such that they can be true or false.

Fair enough. But what about beliefs? Are they not they either true or false? And is that "They're self-defeating!" not precisely what we are wont to say of certain desires? And moreover not just in 'ordinary language'?

Let us deal with desires first. That we often say of, for example, selfishness, that it is self-defeating, I do not deny. But by something's being self-defeating we do not necessarily mean it is self-effacing. For example, suppose I hold that, "When we die all will be lovingly revealed!" Then I might hold that what was (albeit mediately) self-defeating for Faust was his impatience. His desire that all be revealed now defeated his desire to be with his Creator in the Hereafter. But from this it does not follow that his desire that all be revealed now is (directly) self-effacing. Quite the contrary: had he effaced that desire it would not have been fulfilled.

So if we must have a notion of self-defeat which will be a property of ontological kinds other than propositions, let us at least distinguish it from self-effacement. Let us say a desire is self-defeating just in case: the entertaining of it precludes its satisfaction. Whereas by its being self-effacing we will mean: its forfeiture promises its satisfaction.

And similarly, then, might we handle (the possibility of) beliefs having truth-values. That is, one could simply stipulate this possibility away by insisting that though 1) beliefs have no truth-values, 2) propositions do and 3) propositions are the _contents_ of beliefs.- But then the question is simply begged by the notion of a 'content'. And if one elects instead to simply drop (3), along with it she drops all our intuitive associations between beliefs and what, if anything, they might be about, thereby setting herself to task telling a very long story indeed about what residual connection, if any, the two might nonetheless have.

Such a 'naturalized epistemological' story is precisely what I might someday want to tell. But one _need_ do nothing of the sort. One can adopt the conventional wisdom that beliefs have truth-values, that they are therefore candidates for self-defeat, but note once again that the notions of self-defeat and self-effacement are distinct since: the former cites a satisfaction-preclusion condition whereas the latter picks out a satisfaction-enabling one. So, for example, my belief that I believe nothing - or my doubt that I doubt - are self-defeating because they preclude the possibility of the belief being true. But neither are directly self-effacing. Nor under most circumstances are they even mediately self-effacing.

14. Self-Deception:

And, last but not least, self-effacement is akin to, but not quite, self-deception. Let us see why.

Since the seminal pieces by Raphael Demos, "Lying to Oneself" in 1960 [15], John Canfield and Patrick NcNally, "Paradoxes of Self-Deception" in 1961 [16], and Canfield and Don Gustavson, "Self-Deception" in 1962 [17], self-deception has been modelled as everything from a conceptual hoax to a proof for the pervasive existence of multiple mentality. [18] For my own part, I must admit, I find little of interest in this debate. For in the first place I am inclined to reject the unreduced, folk-psychological categories with which the phenomenon has typically been framed. Since, that is, not unlike Hume I am a skeptic about the very existence of 'selves' - since, not unlike Locke with respect to 'persons' I am an instrumentalist about 'selves' - identifying and individuating them is, for me at least, a matter of mere convenience. So in the second place I cannot but view the matter as akin to the aforementioned Hart-Fuller debate, i.e. as substantively vacuous.

For what it comes down to, as I read it, is whether or not one and the same self can, at one and the same time, believe both p and not-p. If we decide she can, then we have to modify our traditional ways of identifying and individuating selves. If we prefer to keep our traditions - and so deny that one can simultaneously believe both p and not-p - then, though there may yet be some phenomenon to be explained, there is no self-deception to be explained, and hence no paradox to be explained.

This is not to deny that, for the instrumentalist as well as the realist, there are devilishly intractable difficulities involved in deciding which way to go here. As indeed there are with the isomorphic Hart-Fuller debate. In fact at the outset of his own attempt at unraveling the self-deception puzzle, Herbert Fingarette puts his finger on this intractability quite nicely. [19] What we want on the one hand, notes he, is an account of self-deception that takes the phenomenon seriously and is true to it. That is, whether aptly named or not, people do seem to do something we have come to call 'self-deceiving', something in significant ways dissimilar to things for which we have other names. For example, they seem to believe (or fail to believe) things for which they seem to lack (or have) the normal epistemic warrants. So there is at least some sense in which they are deceived. And yet there does not appear to be anyone involved in this deception other than themselves. So 'self-deception' certainly seems to be an appropriate enough word for what is going on.

But on the other hand, cautions Fingarette, what we do not want is an analysis of the phenomenon that ends up so compartmentalizing and/or individuating 'selves' and/or 'persons' that the very notions cease to be of any use to us. [20] So, for example, we cannot characterize self-deception as

> some person-stage A rendering some unit of information U erstwhile available to A cognitively unavailable to some continuer of herself B.

For then holding B accountable for her deeds as if she had access to U would do grievous injury to very dear forensic intuitions. Likewise to hold B accountable for A's so rendering of U. And yet exculpating B in this way gives A too much of an incentive to self-deceive, or, assuming A takes an interest in B, at least insufficient disincentive not to self-deceive.

But neither can we characterize self-deception as some person-stage simultaneously believing p and not-p, since to do so is likewise to disqualify her as a rational moral agent - assuming, that is, being ill-ordered she cannot be counted rational.

What is to be noted, however, is that so far, at least, the problems of self-deception and self-effacement seem of a piece, if not identical. So, it might be queried, is one just a sub-species of the other? And if so, which of which?

Let me say, from the outset, that I am by no means convinced that whatever it is that people call 'self-deception' is exhausted - nor even that it is in any wise captured - by the analysis that follows. But if it is, then, I suppose, the story would go something like this:

Jack, though happily married and committed to fidelity, nonetheless wants to have an affair. What stands in his way is the inference from "having an affair" to "committing an infidelity". So he (albeit probably only temporarily) 'obtunds' the inference - or renders it 'cognitively inaccessible', call it what you will - and replaces it with something like, "But of course it really doesn't mean anything!" Similarly, when, in the course of his love-making, and in order to enhance the experience, he needs in turn to obtund this 'meaninglessness', and yet cannot rid his mind of thoughts of his wife, we tells himself, "But surely there's enough love in the love jar for more than one woman!" And so on.

Now just exactly how to characterize this 'obtunding', or 'rendering cognitively inaccessible', is precisely what the self-deception debate is all about. Three questions are posed therein. Is such obtundence psychologically feasible? If so, how is it done? And upon whom, the obtunder or the obtundee, falls the responsibility for having done it? And, sure enough, are these not precisely the questions with which, in the matter of _The Man in the Glass Booth_, we started out?

If this analysis is correct, then self-deception turns out to be a case of _temporary_, cloaked, internal, self-disciplining, contented, strong, symmetrical, mediated self-effacement. That is, what distinguishes it from self-effacement simpliciter is that the exigencies of the game require that one be able, indeed inclined, to toggle back to the original position upon completion of the encounter - in Jack's case the affair - for which it is designed. (Having leapt James' chasm, does one not look back in amazement at having just done the impossible?!)

Or, if one prefers, self-effacement turns out to be a case of _permanent_ self-deception; and what distinguishes it from self-deception simpliciter is that the exigencies of the game require that one be disinclined, indeed incapable, of toggling back to the original position upon completion of the encounter - e.g. moving first in a sequential PD - for which it is designed. But if all this is so, how significant is this distinction?

Significant enough, I think, to present paradoxes for the one (self-effacement) that need not plague the other (self-deception). How so? Because suppose that, to purchase some respite from his tribulations, Goldman had merely imagined himself as Dorff, secure from predation, luxuriating in his officers' quarters, quaffing a fine wine, and so on. Do we not we all daydream? And is there not a kind of temporary suspension of disbelief involved in all such fantasizing, a kind of temporary obtunding of (what one surely knows to be) reality?

So wherein lies the line between Goldman's unproblematic fantasizing and his full-blown self-effacement? Between imagining himself in the circumstances of Dorff and imagining himself to be Dorff? But surely this merely begs the question. At what point does the first become the second? Or should we avoid this difficulty altogether and insist that what is problematic (because loathesome) is even the imagining of oneself in the circumstances of one's value-nemesis? In other words, is there a principled line to be drawn here, or just a Sorites problem that must, accordingly, be handled with a multi-valued logic of acceptability?

To these thorny questions I have no answers. But note that such questions do not arise in the case of full-blown self-effacement. So one way of looking at self-effacement is as a kind of test case - precisely because it is the limiting case - of the logic of self-deception. Well then, perhaps we should settle for that.

8. Resolving the Paradoxes of Self-Effacement

1. *Choices* - *Hers, His, and Mine*:

Most of us are familiar enough with *Sophie's Choice*. For the last seven chapters we have been familiarizing ourselves with Arthur's. That the former is gripping, and wrenching, I trust there will be no dispute. But whether more so or less so than had Sophie been forced instead to choose between one of her children and her *own* life (or *mutatis mutandis*, the man who threw himself on a grenade to save mine), or between her life and her *God* (Joan or Arc), or between her life and her (as it happens mistaken) convictions (*A Man for All Seasons*, Victor Hugo's *Ninety-Three*, and so on), will depend on with just what values our own peculiar experiences have infused us. Is there a mind-independent fact of the matter about which *should* be the more compelling story? Is there something 'wrong' with people who are as (or more) moved by the sight of a dead dog on the highway than pictures of the victims of the Bosnian civil war? Is there something wrong with *me* that I am more moved by Goldman's choice than Sophie's? Or is it, as some might suggest, just one of those - 'gender things'?

Hard to say. But here is another (largely) feminist insight. Masculinist ethicizing veritably _revels_ in paradox. But paradox, in ethics at least, arises only when the _conditions_ for (anything even approaching human morality) are withdrawn. So why not draw instead from _The Man in the Glass Booth_ an answer to the very question I so summarily dismissed at the outset of this enquiry? That was, recall:

> Even [were] it possible for the human mind to perform the kinds of gymnastics which, where the story to be believed, would have to have been the psychological history of the man in the glass booth, _should_ such gymnastics, be they possible or not, ever be necessary?

That is, why not take it as the task of ethics to see to it the conditions under which _both_ Sophie and Arthur faced the choices they faced - "Never Again!" arise?!

That is exactly what I take the task of ethics to be. But my suspicion is the conditions under which Sophie and Arthur faced the choices they faced were themselves the product of people like Dorff self-effacing the very humanity that normally precludes such conditions from ever arising. So now *I* seem to have a choice. Either I can say that:

> given the exigencies of the game called Surviving the Third Reich, it was perfectly reasonable for people like Dorff to have self-effaced their humanity. Given the exigencies of the game called Surviving the Treaty of Versailles it was perfectly reasonable for ordinary Germans to have self-effaced the Weimar Republic and replaced it with the Third Reich. Given the exigencies of ... and so on, back to the moment of Creation. So the tribulations of Sophie and Arthur - don't you see? - are just the inevitable outcome of human rationality!

Or I can say that:

> notwithstanding the exigencies of the Third Reich, the Treaty of Versailles, or what have you, one's humanity, or one's democracy, or one's what have you, are things *not* to be self-effaced. And so it was a *mistake* for Dorff, or ordinary Germans, or whomever, to have done so!

But to avoid such mistakes in the future, it seems to me we need to begin by identifying them. To have a notion of a 'mistake' we first need a norm by which to measure it. And to have any confidence in our norm we need to test it against the limiting case. And so we find ourselves once again with The Man in the Glass Booth.

But, counters my detractor, now I am begging the question. To have confidence in a norm we must, true enough, test it against the limiting case. But the question, recall, was whether Sophie's and Arthur's choices fell within or without that limit. I say yeah, she nay. So how do we resolve the issue?

Or, to put it another way, we are agreed Leroy's choice falls within the limit. We are agreed the norm is to be interpolated from our intuitions regarding choices the likes of Leroy's. Where we disagree, it seems, is over whether we can extend that norm to cover the choices of Sophie and Arthur. So, she might say to me, confine the extrapolation of the norm to the range of cases from which it was interpolated in the first place, and our dispute is at an end.

Or, to put it a third way, we are agreed like cases are to be treated alike. (Without that axiom all thought worthy of the name comes to an abrupt end!) What we are disputing is whether or not Sophie's and Arthur's choices _are_ like Leroy's. For in a world like Leroy's - bizarre as the Only may be - with instrumental rationality seems an appropriate way to respond. But in a world like Auschwitz?! There instrumental rationality is a masculinist philosopher's pipe-dream!

And, I must admit, put this third way her point begins to take its toll, even on _my_ phallologocentric soul. So perhaps _The Man in the Glass Booth_, notwithstanding its Bachian charm, is just too penumbral a case to which to attempt to apply any of our findings, whatever they might turn out to be. And perhaps _that_ is why our sense that

> beyond the crimes against humanity (as standardly understood) that may have transpired here, and our search for someone to hold accountable for them, there is another crime that has been committed here

is so only-cloudily felt, as so only-cloudily felt is

> the obscen[ity] and profound unnatural[ness] of the patricide of Oedipus, the regicide of MacBeth, or the deicide of the Passion.

Perhaps, instead of a puzzle case in ethics, *The Man in the Glass Booth* belongs in the canon of *religious* myths, the *forte* (and value) of which being precisely their impenetrability, their very mystery. After all, can the obscenity and unnaturalness of 'autocide' - if such the case of *The Man in the Glass Booth* be - be 'explained' any more than that of patricide (and incest), regicide, and deicide?

Or at least that is one way to go. Here is another. The obscenity and unnaturalness of autocide, patricide (and incest), regicide and deicide *can* be explained. Moreover they can be explained by a common explanatory schema. What it is to be human, at least in part, is not just to (take oneself to) be going somewhere, but also to (take oneself to) be coming from somewhere. (Nor need I be saying anything terribly obscure or 'continental' here. Rather all I need be saying, or at least rhetorically asking, is: How could it be otherwise with an organism whose survival and wellbeing are so dependent on the pursuit of long-range projects?!) One 'comes from' a father, a mother, a kingdom and/or a God. Likewise does one come from a (logically and/or temporally) prior self. So killing one's father, king, God and/or prior self, is as much a disruption to one's sense of 'location' as is intercourse with one's mother a 'sullying' of the womb from which one once emerged.

Such inferences are not - nor need they be - rational, if by 'rational' is meant merely 'reasonable'. But by being 'associative' or 'symbolic' they might still be 'rational' in the instrumentalist sense of the term. Why do we make them? Either because respect for womb, sire, authority, God and/or prior self was advantageous and/or because, cf. the Principle of Sufficient Reason, such respect was not _dis_advantageous. Just _why_ respect for womb, sire, authority and God might be advantageous (or not disadvantageous) there are stories aplenty. As there are for respect for one's prior self.

Of course I need not be committed to this particular meta-story. No doubt there are other, perhaps more plausible, candidates. But my point is that _that_ we have intuitions which are only-cloudily felt may itself be explicable. So _that_ we may feel we are profaning sacred ground in presuming to extrapolate our norms onto narratives like Oedipus, MacBeth, the Passion and _The Man in the Glass Booth_, does not in itself bespeak the inappropriateness of doing so. What it says, and all it says, is - remove your sandals and proceed with heightened reverence!

2. An Interim Summary:

Let us summarize what we have found so far.

We began by observing that what The Man in the Glass Booth challenges us to think about is not so much what, if anything, we owe to each other, but rather what, if anything, we owe to ourselves. And, not surprisingly, we concluded that we must owe ourselves something since, our intuitions tell us, there are more defensible ways than others of projecting ourselves and our values into the world.

Of course a complete theory of defensible self-projection would be far too ambitious a project for one library much less one volume. But most of the ways in which we project ourselves are relatively unproblematic. At t1 I am hungry and there is an apple on the table. I can make it the case that at t2 there will still be an apple and I will still be hungry, or the apple will be gone and I will not be hungry anymore. Deciding between the two is then a simple matter of consulting my current preferences between these two anticipated states of affairs.

Of course it is trivially true that satisfying virtually any desire will simultaneously change something about me. But what The Man in the Glass Booth is designed to test for is whether or not there is anything about myself I cannot defensibly change, notwithstanding that undergoing that change might be the only way of satisfying the desires I now have. In other words, are there any constraints on the defensibility of self-projection other than the instrumental fit between present and projected selves? Does our niggling discomfort with the man in the glass booth tell us the answer is yes? Or can this discomfort be accommodated by simply postulating a second-order desire, a desire not to violate some (perhaps arithmetic or geometric) limit on just how radically we can allow our first-order desires to change?

3. Our Options:

Now it seems to me there are four conflicting intuitions about the purpose or 'meaning' of existence, and hence, parasitically, of life. The first - as old as Plotinus and as recently rearticulated under the rubric of axiarchism by John Leslie [1] - is that

1) there is a mind-independent fact of the matter about what is valuable about existence; and that -

although this is where Leslie jumps ship - therefore, insofar as

2) (at least part of) what that is can only be instantiated by a psychology,
3) what is incumbent upon us is to bring our psychologies in line with the exigencies of that instantiation.

So, if that value is, say, pleasure or happiness or fun or whatever, then clearly we should operate on Beatrice and, just as clearly, we should reach for the Nozickian cord. Note, however, that if the designated value is 'freedom' or 'integrity' or 'fulfillment' or any of their cognates, then the theory is no longer of this category, since these notions - unlike the axiarchic notion of value - are subtantively contentless.

We reject this view - if need be, out of hand!

Our second option, as already noted, is some kind of mathematical sufficiency condition, usually (cf. Parfit et al) on psychological similitude, or contiguity, or some combination of the two. On this view what is wrong with the man in the glass booth is just that he allowed himself to change too radically and/or quickly. Similarly with the gay military recruit. And similarly not so with Leroy.

But this will not do either. For in the first place the conversion from SM to CM is surely a radical one; and, were pharmaceutical means available, rapid as well. And yet no one would deny it is a defensible one. And in the second place we often make incremental conversions in the full knowledge that - indeed precisely because we know - they, in turn, will trigger further incremental conversions ... and so on ... which will result in a radically different psychological state of affairs. But we often forgo even the first step precisely _because_ we do not like where it will eventually take us.

For example, I was once offered - and I declined - a certain career which, I knew full well, would certainly not have changed me overnight. Moreover I would have found it deeply satisfying. But I was also convinced it would in _time_ make me the kind of person I preferred not to become, precisely _because_ it would make me the kind of person who would eventually enjoy things I preferred not to ever enjoy. And similarly, we might say, of Goldman and G-D.

That the defensibility of self-modification is to be mathematized in some such way is precisely what non reductionists and/or essentialists about 'what matters in survival' deny. [2] What is at issue, say they, is not how many (or what proportion) of 'units' of similitude or contiguity can be defensively violated. Rather it is a matter of maintaining the integrity of some privileged unit, be it the soul (primitive Christianity), the terms of some Covenant (primitive Judaism), or what have you. On the essentialist view what may be wrong with forfeiting one's homosexuality in order to pursue a career in the military - but not one's desperation in order to 'meet' someone - is that one's sexuality - but not her intensity - is a sine qua non of her being who she is.

But essentialism, it seems, is incompatible with the kind of contractarianism we have been running with here. For Hobbes (implicitly) and Gauthier (explicitly), rationality - by which is meant maximizing on the satisfaction of preferences - is to be understood as utterly indifferent to the substantive content of those preferences. Why? Because Gauthier is especially (and understandably) keen that his theory of rationality be given as wide a berth of application as possible. If, for example, one were to stipulate - as, for one, does Rawls [3] - that one must have some minimal set of peculiarly human values to start with before she can qualify even as a candidate for the kind of rationality covered by the theory, one runs the risk of ruling the psychopath out of the domain of discourse. And yet it is precisely the psychopath to whose rationality Gauthier wants to appeal to coax him to 'change his mind'. In other words, by having it adopt an essentialist stance, one's theory of rationality has little (if any) work left to do. All (or most) of the work done by the theory is done instead by (and within) the debate internal to the theory over which properties are to be designated 'essential'.

And so our fourth, and final, option is to advance instead a theory of - prudence.

4. The MacIntosh Gloss on Prudence Theories:

By a 'prudential' theory, in this context at least, is meant any view according to which the defensibility of an agent's self-modification is exclusively a function of its instrumentality towards the satisfaction of that agent's existential desires. So neither naturalism nor essentialism are prudential theories because the constraints they place on one's self-modifications are independent of what she in fact desires. Rather they make reference to, respectively, what - *qua* instance of her kind, or *qua* her own peculiarities - she *should* desire. Neither, therefore, is 'mathematicism' a prudential theory. For even though, unlike naturalism and essentialism, it espouses neutrality on issues of erotic content, the relation (between antecedent and subsequent selves) with which it concerns itself is not degrees of likelihood of satisfaction, but rather degrees of similitude and/or contiguity.

Now by its giving (what prudence theorists see as) proper weight to 'existential' desires, they do not mean the desires to be thus weighted need necessarily be current. As we will see, some prudence theorists - Duncan MacIntosh among them - might suppose (or argue) so. But if they thought otherwise they would remain prudence theorists nonetheless.

Foremost among those who have been thinking long and hard about prudence is, as I say, Duncan MacIntosh, for whom the key issue is precisely whether the desires that drive rational, prudential self-modification need be concurrent with the choice, or whether they might as well be merely anticipated and/or remembered. [4] MacIntosh wonders whether we might accommodate our intuitions about prudence by supposing we perform a kind of (albeit weighted or 'isobaric') utilitarian calculus across remembered, current, and anticipated stages of ourselves. And, he concludes, this will not do. Why? Because, he asks rhetorically, What could possibly possess me to compromise on the satisfaction of any of my current desires save that I currently have a second-order, a past-regarding and/or a future-regarding desire that the satisfaction of the desires of remembered or anticipated stages of myself be awarded (albeit weighted or isobaric) consideration?! But if this is the case, as he points out, then there is no reason to look beyond my current desires. And if it is not the case - if, that is, I currently have no such second-order, past-regarding and/or future-regarding desires - then it would be nonsensical to sacrifice the satisfaction of any of my current ones.

Or, put yet another way, even if we suppose past and future states of oneself are to be regarded as 'other people', and that we are impartial (albeit weighted or isobaric) utilitarians, we would not currently be utilitarians unless we currently cared about (i.e. had preferences regarding) these other people.

And, put this way, it is hard to disagree. The (albeit only putative) difficulty with his position arises, acknowledges MacIntosh, in (what he calls) 'paradoxical choice situations', or PCSs. [5] (It may be noted that MacIntosh finds PCS's problematic in precisely the way Gregory Kavka found Special Deterrence Situations, or SDSs [6]; which is not surprising given that SDSs are, MacIntosh notes, merely instances of PCSs.

[7] Others include the PD, the TP and the NP. [8]) So what is "news" about PC's? asks MacIntosh. Just that

> it [can be] rational to acquire a preference for x not because x is causally needed for or logically part of something else you already want, y, but because preferring x is needed to get you y. [9]

MacIntosh considers three challenges to the rationality of changing one's preferences in a PCS. The first is particularly germaine to our own enquiry, since the scenario it cites is precisely that of the regulars down at the Only requiring of Leroy's desperation that it be strongly self-effaced. Likewise, then, does Bob Bright asks us to

> 1) suppose that to cause what you prefer, you must disprefer it - "I'll give you what you now love if you will hate it by when I give it to you". In acquiring the new preference, you would also be arranging that it not be satisfied; and this seems to violate the rational obligation to maximize one's expected utility, to satisfy one's preferences. So you should not change. [10]

Likewise does David Zimmerman echo worries we have ourselves been considering when he insists that

> 2) Satisfaction involves not just a condition's obtaining, but also one's preferring it. Thus, one cannot cause a preference's satisfaction by ceasing to have it by when its target condition obtains, for then it no longer exists to be satisfied. Thus losing a preference cannot be a means to its satisfaction. So [once again] you should not change. [11]

And, finally, some of the Parfitian moves we have been considering would seem to dictate that

> 3) rational agents must satisfy their preferences. But agents [just] are their current preferences. If something has different preferences from those you now have, it cannot be you. Thus you cannot cause your preferences' satisfaction by changing them; you would no longer exist, so nothing could then count as satisfying your preferences. So [once again] you should not change. [12]

MacIntosh's answer to (1) is to point out that it presupposes that "rational agents advance their current and future preferences" [13]; whereas his view, recall - and did it not seem reasonable? - is they advance only their current ones. Against (2) he counters that

> if i) utility is what you get when a condition obtains which satisfies a preference you have when it obtains, [then] ii) you [could not] rationally have and advance goals whose attainment could not raise your utility. [Well], iii) the dead have neither preferences nor utility. But surely iv) you can rationally have and advance goals involving your own death (e.g. providing from your family when you die). [14]

So (i) must be false, as must (ii). And, finally, (3) supposes what clearly cannot be the case, namely that "every character change is [a] suicide." [15]

One could, I suspect, mount an argument against (iii) and/or (iv), thus salvaging (i) and (ii), and with it (2). And perhaps one could advance a more plausible version of (3). But MacIntosh himself is less concerned with conclusively refuting these three objections than with exploring [16] which of what he takes to be the "three conflict[ing] theories of practical reason" [17] -

what I have called 'prudence theories' - will best meet them. On the first view,

A) it is rational to cause whatever one prefers. So it is rational to change a want to satisfy it. On the second,

B) it is only rational to cause the utility of attaining ends one will want when attained, and one must have attainable ends. So it is only rational to change a want where that will cause a higher utility over one's life given all the preferences one will ever have, not just those one had when contemplating the change. And on the third,

C) it is only rational to cause the utility of attaining ends one currently wants and will still want when attained. So it is only rational to change a want if that will raise one's utility by the preferences one now has. [18]

Now MacIntosh's proposal is that "a choice is rational only if it maximizes on preferences [that are] rational to have." And

> a currently held set of preferences, P, is rational [to have, in turn,] only if there is no other set, P*, the having of which is more likely than having P to cause P's target conditions in the order preferred in P. So a current preference set is rational just if "self-maximizing", [that is] if having it maximizes by its own measure compared with having any other. Thus we apply the maximization test not just to the choice of means to ends, but also to the choice of ends (given current ends). [19]

In other words, counsels MacIntosh, if the most likely way to get what I want is to no longer want it, then I should cause myself to no longer want it. That I will no longer want it when I get it is, on his view, utterly irrelevant. If it is relevant, then, MacIntosh instructs us to ask, to whom is it relevant? To me? Then, quite obviously, what I must have wanted was to have it while wanting it. But in that case no longer wanting it was clearly not the most likely way of getting what I wanted. In fact it was a sure way of not getting what I wanted.

But, counters MacIntosh's detractor, "If preferences are only rational if self-maximizing, surely this holds for prospective preferences too." [20] So consider MacIntosh's own example.* Suppose that, "having tired of defending [my marbles] ... I offer to give you 5 ... just if you come to hate marbles." [21] But then,

> i) if having the preference for 0 marbles will make you get 5, it is not self-maximizing; it prevents its own target. Thus surely
> ii) a rational agent will adopt preferences maximizing on his originals except where that would not maximize on the new ones. [22]

Not so, counters MacIntosh. "Preference revision," he insists, "is rational even if it frustrates foreseen preferences." True, he concedes,

> now I want to spend all my money on movies, [even though I know that] when I am fifty I shall wish I had saved it and bought a house ... But prudence cannot always be rational. For if it is, one must treat one's foreseen preferences as if they are, like one's current ones, relevant to current choices.

Treating foreseen preferences like current ones amounts to treating them like simultaneous ones. But that means simultaneously preferring movies to a house and a house to movies. "But," MacIntosh reminds us, "one cannot maximize on ill-ordered preferences." [23] So, he concludes, the only way to parry objection (1) is to adopt theory (A).

As, claims he, is the only way to parry objection (2). Why? Because

> the notion that rationally to prefer x one must also prefer still to prefer x by when it obtains has [unacceptably] absurd consequences. First, if one had both some preference and the preference to keep it, one could not get into a PCS [in the first place]. But PCS's are possible. [And] second, if every preference for an outcome also involves a preference still to prefer it upon the former's satisfaction, preferences for events after one's death would be irrational. [24]

Which, clearly, they are not.

Wherein lies the second objector's mistake? It lies, claims MacIntosh, in his conflation of "the rational duty to maximize expected utility [with] the duty to maximize utility, [the latter being] a condition's obtaining while preferred." [25] But, MacIntosh points out, "my utility does not rise from the obtaining of a condition I no longer prefer - especially if I am dead." [26] So

> preference satisfaction and utility [cannot be] equivalent. Having utility entails that a preference is satisfied, but that a preference is satisfied does not itself entail having utility. Thus preferences can be satisfied by the obtaining of their target conditions even if the preferences or their holders have expired by then. It is just that satisfaction then does not yield utility. [27]

Of course normally, MacIntosh concedes, one can expect to get utility from the satisfaction of one's preferences. Some PCSs, it seems, are exceptions. "But," says he, "no matter. [For] one's only rational obligation is to seek [the] satisfaction [of one's preference], not [or at least not necessarily] utility from its satisfaction." [28]

There can be little doubt theory (A) does a better job than either of (B) and (C) at keeping the logic of decision 'neat and tidy' in the face of objections like (1). And it does a better - and certainly more ennobling - job than either of (B) and (C) of making sense of our beyond-the-veil-regarding concerns. But MacIntosh wants to get a lot more mileage out of these advantages than I think they warrant. What he wants is to defend what he calls the 'received' theory of rational choice. The received theory he calls MIEU, according to which we are to:

> maximize one's individual expected utility, maximize the probability and current preferability of conditions given choices.

And he wants to defend it against two utility-based pretenders, MEIU:

> maximize one's expectation of individual utility, maximize the probability and concurrent preferability of conditions,

and MEIUCP:

> maximize one's expectation of individual utility by the rigid measure of current preferences, maximize the probability and concurrent preferability of conditions by _current_ preferences. [29]

What distinguishes these pretenders from each other, then, is just

> what utility is worth caring about. MEIU obliges prudence: aim at satisfying current _and_ future preferences concurrently with having them. It sees no rational difference between them, so one must concurrently maximize on both. This can mean sacrificing the concurrent satisfaction of a current preference to that of a future one. [30]

So with regard to the considerability of current and future selves, MEIU is egalitarian; on distributive justice it is utilitarian; and as to virtue ethics it is stoicist. That is, since "having unsatisfiable preferences reduces one's possible utility, one should only prefer what one can get." [31] According to MEIUCP, by contrast,

> [there is no] utility worth wanting now from satisfying preferences one does not _yet_ have. [So whereas] MEIU requires one to treat one's future preferences as if they were current, MEIUCP [demands] that one _not_. [32]

So, in short, an agent employing MIEU will be prudent because she _now_ cares about the future. One employing MEIU will be prudent because she _will_ care about the future in the future. And one employing MEIUCP will be prudent because she cares that she _will_ care about what she cares about now.

Now then, assuming MIEU, MEIU and MEIUCP exhaust the field, "what decides the correct theory?" [33] All three cover normal cases of rational choice. For what it is worth MIEU seems 'cleaner'. And it seems to do best at covering beyond-the-pale-regarding concerns. But are cleanliness, if such a thing there be, and user-friendliness to the soon-to-be-dead, sufficiently compelling on their own? MIEU, with its emphasis on satisfaction over utility, is less hedonistic, and so (perhaps) more ennobling. It can make sense of, for example, the deaths of Socrates and Thomas More in a way the other two cannot. But 'nobility' is already either a moral category or an aesthetic one; and it would hardly do to allow either a place in adjudicating the very theory, i.e. of rationality, with which we hope in turn to adjudicate our moral and aesthetic adjudications.

Moreover, whichever we decide handles PCSs best, the other two will accuse us of begging the question. For what counts as handling PCSs best will depend on which, for independent reasons, we already think is best. Of course The Man in the Glass Booth is a PCS with a vengeance; so a fortiori we cannot test these theories against our intuitions in that case, since, that is, it is precisely in the case of The Man in the Glass Booth that our intuitions are most singularly lacking. Our sole intuition, if such it be, is there is something not quite right about what Goldman did. But that is hardly the wherewithall for a theory of rationality.

MacIntosh takes himself to have a knock-down argument for MIEU from the analyticity of instrumental rationality. Here are his three axioms of means-ends rationality:

1) when a rational person chooses between x and y, if he prefers x he must choose x;
2) given a choice between actions likely and unlikely to cause what he prefers, he must choose the former; and
3) when choosing among actions different in the probability and desirability of their consequences, one must combine (1) and (2) - he may only take expensive risks for strongly preferred conditions, refuse inexpensive risks only for weakly preferred ones, and so on.

So any further constraints on instrumental rationality must be deduced from (1), (2), and/or (3), and/or from beliefs about the circumstances of choice. [34] But, claims Mac-Intosh, "agents who choose by MEIU or MEIUCP will sometimes violate" (2). [35] So to be rational just is to MIEU. [36]

5. Reservations:

So why are we not convinced? For the following reason:

The 'received' theory of instrumental rationality, MIEU, was constructed to explain means-ends reasoning. Prudential considerations challenge that explanation. And, let us suppose, MIEU meets that challenge. PCSs then challenge the meeting of that challenge. And, once again, MIEU meets *that* challenge. But even if we suppose that the limiting case of a PCS, a.k.a. PCS-with-a-vengeance, a.k.a. *The Man in the Glass Booth*, challenges the meeting of *that* challenge, and MIEU meets *it*, we must still ask, At what *price* does it do so?

At the price of postulating preferences so highly ordered that no being of our earthly acquaintance could possibly entertain them! For example:

To explain why I refused the very lucrative career-path that was once offered to me - assuming, of course, that refusal was rational - MIEU insists that at the time I must have entertained a preference that no future stage of myself should become the kind of person who would enjoy doing the kinds of things that someone who had done that particular job for some time would inevitably come to enjoy. But if I am to postulate my having such a desire, surely it is not unreasonable to ask why I should have had it in the first place. After all, if I did not take the job, with all its perks and pleasures, someone else surely would. No doubt someone did. So the desire I must have had is not that things such as the job would involve would not be done - or if done then at least not enjoyed - but rather that no continuer of me should be the one doing them and/or enjoying doing them.

But, once again, why should I have that desire? After all, my continuer would certainly enjoy doing them. And he would certainly not suffer the moral qualms about doing them and/or enjoying doing them that I apparently would. And even if he remembered having had the qualms I now have, he would dismiss those qualms, and along with them the antecedent self having them, in much the way I now dismiss the qualms of selves, and those selves themselves, antecedent to me. In short, to render my refusal rational we need to postulate not just some desire I could have had at the time that would have dictated the refusal, but one which makes some sense. And for much of what we do, especially in cases involving PCSs, this is no small task.

Note that the objection here is not to the very idea of postulating desires. I appreciate that most of our desires are probably dispositional. For example:

"Would you like to go sky-diving?"

"Ah ... sure."

"Had you wanted to go sky-diving before you were asked?"

"Well, no, not that I know of."

Nor is the objection to the idea of postulating desires that are very probably unentertainable in any conscious way.

"Do you prefer to prefer to prefer to prefer to prefer Coke over Pepsi, or just to prefer to prefer it?"

"Hmmm ... Damned if I know!"

Nor is it even to the idea of postulating desires that may not even be dispositional. Asked of that Saharan mother stumbling barefoot across the desert, her one surviving child clutched fly-covered to her own shrivelled bossom:

"What do you want?"

"To get to where there's food and water!"

"Yes, but don't you also want your life to go, all things considered, as well as possible?"

"Please, mister, do you know where there's food and water?"

Rather the objection is that the desires that must be postulated of us by MIEU in order to render our behaviour rational, especially in PCSs, are so abstract and pristine we must be more akin to gods than human beings! But human beings are not gods. Human beings are organisms. So, as we are about to see, perhaps the idea that we perform some kind of (albeit weighted or isobaric) utilitarian calculus across past, present, and future stages of ourselves (MEIU), coupled with some kind of naturalist-essentialist and/or mathematical relational constraint between them, is closer to the truth after all, notwithstanding that it may violate the second of MacIntosh's axioms!

6. The 'Meta-' Move:

MacIntosh's mistake, I submit, lies in his failing to see that our algorithm - or more accurately algorithms - for instrumental rationality are themselves the product of forces which are not themselves rational because they are not themselves erotic. That is, what we want, and how we go about getting it, are ultimately determined by which wants, algorithms, and circumstances are jointly responsible for our continuing to be here. In a world in which a) I want to kill myself, b) I employ standard MIEU to do it, and c) the laws of physics are much as they are, I will probably succeed, and so fail to continue to be here. Vary either (b) or (c) and I may very well fail (in my project), and so succeed (in nature's). Vary (a) and either of (b) and (c) and I may fail, and so fail. And so on.

This is not to say I should want to continue to be here. Nature's projects - if such a metaphor is allowed - are not necessarily ours. But even confining ourselves to our projects, we would be well-advised to tailor our algorithms to the circumstances in which we are embedded.

This, then, is the meta-move. But what does it have to say about PCSs? PCSs, recall, are circumstances such that the satisfaction of our wants are best served by revising them. MIEU is an algorithm that allows us to go ahead and do so. MEIU and MEIUCP are algortihms that do not. They refuse to do so because they assume we have another want which we want satisfied more than the first want, that being the want to want the want that is concurrently being satisfied. MacIntosh points out this is certainly not always the case - e.g. we do seem to have beyond-the-pale-regarding wants. And where it is the case MIEU will refuse to countenance the revision. Which, he points out, is just as it should be.

What the meta-move allows us to see, however, is that it may be an exigency of our continuing to be here that the algorithm we employ for processing our wants may be at odds with the algorithm best suited to satisfying them. That is, it may be that MIEU is, as MacIntosh believes, best suited to satisfying our wants. But it could at the same time be that M*EI*U, or some as-yet-unexamined alternative, is best suited to perpetuating our existence. So this alternative to MIEU, whatever it may be, occupies the algorithm-space, it goes unnoticed - probably because in normal, non-PCSs, it takes identical inputs and produces identical outputs - and it only reveals itself to us, much to our surprise and consternation, when we are confronted with certain peculiar kinds of prudential considerations, namely PCSs.

Having wedded his mind to MIEU being the preferred algorithm for *any* project - be it ours or nature's - MacIntosh acknowledges the abberant output data, and then tries to force MIEU to produce it. What he fails to consider, however, is that though MIEU may be, as he says, the preferred algorithm for any project, it hardly follows that it is therefore likewise the preferred algorithm for an agent (whose 'project' is to produce a certain *behaviour* in an organism) to load *into* that organism. That is, it may be MIEU-rational for *me* to load into *you* an algorithm that it is *non*-MIEU-rational for you to load into yourself; just as, recall, it was rational for me to try to make a soldier out of you notwithstanding it may not be rational for you to be made into one.

As we have seen, it may be MIEU-rational for a would-be suicide to adopt MIEU to ensure his own death. But if you and I are trying to prevent his death, and were, for whatever reason, unable to alter either his desire for death or the availability of the material means to affect it, but _were_ able to have some say in his algorithm, some non-MIEU might very _well_ be the algorithm with which we might want to saddle him, precisely _because_ it will not process beyond-the-pale-regarding concerns. So, to return to our meta-considerations, the question we need to ask ourselves is this: Given

a) _nature_'s 'project' for us, whatever it may be, given

b) i) the preponderance of ourself-held projects within that project that require of us that we be prudential with respect to our _own_ projects, ii) the relative rarity of sub-projects within those projects that involve PCSs, and iii) the relative rarity still of limiting-case PCSs like _The Man in the Glass Booth_, and given

c) that even as far back as Meditation VI cognitive scientists have known that algorithmic impeccability is always in some wise sacrificed to reduce hardware-burden and response-time [37], is MIEU or some non-MIEU the algorithm we should expect? True, there is no particular reason to expect some non-MIEU. But then neither is there any particular reason to expect MIEU. Certainly no a priori reasoning will settle the issue. It is a purely empirical matter, to be settled, if at all, if ever and whenever we have sufficient data about the kind of organism we are, and the kind of world we occupy.

So, similarly with Goldman. The question is not whether Goldman did the right thing according to this or that conception of rationality. Rather it is: What kind of an organism is a Goldman? With what algorithm had he been programmed? Was that algorithm suitable to the purposes and circumstances for which he was 'designed'? Was that algorithm at odds with the purposes and circumstances for which *Goldman* might have designed himself?

All that said - or, rather, asked - note, however, that it will not do to say that, "Goldman did what he did because, after all, that's just the way he was designed, and that's all there is to say on the subject!", any more than it would do to say that, "He did what he did because, after all, that's just the way he designed himself, and *that*'s all there is to say on the subject!" It will not do because we decided, recall, to entertain the possibility that, even on his own terms, Goldman might have made a mistake. And so it will not do because, we assume, one of our design features is an algorithm-self-monitoring module. That is, we are reflective beings. And one of the inputs for reflecting on ourselves is reflecting on the self-reflections of others. [38] So we can, and do, to some degree at least, revise our algorithms more to our liking; which may, for all we know - or need care! - frustrate the hell out of our Maker!

7. What Decides Between the 'Correct' Theory and the Meta-Theory?

Paradox is the child of 'levels' confusion. Paradox is a species of equivocation. Index by level and the equivocation disappears. As does the paradox along with it. So the question should never be, "What is rational?" "What is rational?" is not a well-formed formula. Well-formedness requires, "What is rational at [some specified] level of analysis?"

So, what can we say about Goldman? Perhaps just this:

Goldman was faced with a choice between becoming what he himself would prefer did not come into being at all and ceasing to be entirely. Given that we have reason to believe _he_ had reasons to prefer the latter to the former, and that he was fully rational at the time, we conclude that he rationally _ought_ to have chosen to have ceased to be entirely. But, apparently, he did not. So now we have a very unpalatable choice. Either we revise our reading of his preferences, or else we withdraw our assumption that he was fully rational at the time of the self-effacement.

Why unpalatable? Because if we choose the former we commit ourselves, or at least Goldman, to the view that some life - no matter how low in _expected_ utility given his current preferences - has greater expected utility for him than no life at all. So according to MacIntosh, if _ex hypthesi_

1) Goldman were fully rational, then
2) he would have to have been MIEU-rational. So by his behaviour he betrayed that
3) he, Goldman, preferred life for G-D over no life at all notwithstanding the irrelevance (to what utility Goldman can expect to reap) of the very _high_ utility that Goldman can expect G-D to reap _and_, at the same time,

4) notwithstanding the *relevance* of Goldman's current dispreference that G-D reap such high utility.

In other words, Goldman must have been a veritable *monster*!

And if we choose the latter - if, that is, we deny Goldman could have been fully rational - on what grounds do we presume to do so? On the grounds that, from the vantage of our very comfortable and secure ivory towers no rational being could suffer the likes of G-D to live, even at the cost of suffering oneself to die? But to presume to say *that* reveals *us* as monsters!

So MacIntosh, it seems, is caught between a rock and a hard place by *The Man in the Glass Booth*. *The Man in the Glass Booth* reveals MIEU for exactly what it is: a theory about what it is to be rational that can only be espoused by people who, in judging the likes of Goldman to be either monstrous or irrational, reveal their own monstrous insensitivity to the fact that being rational is *not* just a matter of maximizing on the satisfaction of current preferences, but rather doing the best one can in that regard while labouring under the burden of a project the author (or Author) of which may be entirely other (or Other).

The meta-theory, by contrast, acknowledges that 'other' (if not that 'Other'), and along with it that 'burden', and so is in a position to exculpate Goldman - if exculpation he needs - without revising his preferences or downgrading his rationality. It does so by incorporating into its theory of human rationality the acknowledgement that as human beings (maximizers on the satisfaction of current preferences, perhaps?) we are nonetheless constantly doing battle with our being first and foremost beings, beings created and sustained by algorithms created and sustained in turn by a process (or Will) which is, both logically and ontologically, prior to our being human.

Just what that process might be, I need not pronounce. I am myself an atheist. So, I suppose, for me that other is natural selection. But I share my life with a devout theist, for whom that other is an Other. And I share a history with a people for whom it is a Covenant. But for me (qua atheist and lapsed Jew) to presume that theism and Judaism are merely beliefs, and as such can be simply processed along with all the other inputs into these poor souls' network of preferences, is not just to trivialize the force of these beliefs, it is to entirely miss their point. The point (or question) is: How did those beliefs get there? For most theists and Jews of my acquaintance would sorely prefer to be rid of the burdens of their faith and their covenants. That they cannot, it seems to me, is testimony to be taken, if not as gospel, then at least seriously.

As seriously and cautiously as we should take Goldman's having broken covenant as evidence that the 'other' is natural selection rather than God, or as evidence that the god of natural selection is (at least) stronger than that of Israel. For *ex hypothesi* it is entirely possible - is it not? - that Goldman broke covenant not because the force of natural selection pulled stronger at him than could the bonds of covenant, but rather because he judged - rightly so I would argue - the co-Signator had already broken covenant with Goldman!

8. How to Be - and Suffer Being - One's Own Other:

I said at the outset of this enquiry that Hiller's AFT made-for-TV adaptation took liberties with Shaw's 1967 novel. In fact Shaw dissociated himself from Hiller's production - somehow, apparently, the copyright got away from him - precisely because it failed to incorporate what was for Shaw the story's core element. Shaw's original story goes something like this:

As his ordeal was coming to an end, Goldman realized he was likely to survive it; and that whether or not Dorff himself survived it, people _like_ Dorff most certainly would. Furthermore, for the very reasons cited for Dorff self-effacing to Goldman, they _would_ self-efface, if not to Goldman then, more generally, to ordinary German citizens who had _not_ self-effaced their humanity. In other words, they would _deny_ the Shoah had ever happened, or at least that they had had any part in it. And since they would have first and foremost convinced themselves, there would be no grounds upon which the rest of the world might remain _un_convinced. Moreover, Goldman realized, their self-effacement would prove credible precisely because the Dorffs of the Third Reich were so _in_credible!

But Goldman was determined the world should _not_ forget that the Third Reich produced monsters the likes of Dorff. So since the Dorffs of the Third Reich would most certainly self-efface to the likes of Goldman, Goldman determined there would be at least one Goldman who had self-effaced to the likes of Dorff so there would be a living, breathing, and _un_-self-effacing testimonial - a memorial, if you like - to the likes of the _true_ Dorff. But since it would take time for the world to forget - and time too to self-efface to Dorff - he masqueraded as himself (if such a locution makes sense), while all the while cultivating Dorff within himself. And in the meantime the residual Goldman within him planted a trail by which Mossad might eventually catch up with the Dorff within him. Thus Goldman's finest hour was his Dorff

> standing proud and erect, recounting in bone-chilling detail, and with great relish, the indignities that he, Adolf Karl Dorff, once had the privilege of visiting upon the members of their families, [and his] even remembering some of them by name.

As was the dental records revealing his true identity his ultimate defeat.

But - or so we might console Goldman - was it entirely a defeat? Imagine the likes of Dorff,

> standing proud and erect, recounting in bone-chilling detail, and with great relish, the indignities that he, Adolf Karl Dorff, once had the privilege of visiting upon the members of their families ...

only to discover that he is in fact Goldman. Could there be a finer moment of divine justice?!

The upshot of this story is we need not suppose the other must be an Other, and so devolve in our analysis of the phenonemon to a religiosity alien (and alienating) to our more secular interlocutors. Rather we can ourselves be the authors of the moles or viruses against which we find ourselves bridling. And this is the sense in which one's faith (Christianity) or one's 'covenants made' (Judaism) cannot be captured by citing them as 'mere' beliefs.

That they are beliefs is not in dispute. What is in dispute is whether they can be defensibly revised. On the MacIntosh view they can be just in case one currently prefers to revise them. But, for example, there would be no point in replacing SM with CM if, having elicited one's co-player's cooperation in a sequential PD, the CMer were then free to prefer to revise her cooperative disposition. But, as we have just seen, sometimes the replacement psychology is epiphenomenally frustrated. Usually this frustration is an unanticipated consequence of the self-effacement, as in:

> I thought that when I allowed myself to be dropped off in Pittsburg that - in addition to the

> sufficient condition of my having done so, that being needing to get back into shape - after a couple days I'd also be glad of having done so. But now that I'm still only halfway up Tuscarora Mountain I see I was altogether too optimistic.

Sometimes it is anticipated but unintended, as in:

> Look, I know that when I'm halfway up Tuscarora I'm going to regret like hell ever setting out. And I'd rather it not be that I'll feel that way at the time. But that's part of the price I'm willing to pay to get rid of these disgusting twenty pounds.

And sometimes it is anticipated and intended, as would be the case if we supposed:

> Goldman knew that, once faced with Eichmann's fate, G-D-G-D would sorely love to have self-effaced back to G-D-G-D-G if he could. But having embraced a Psychological Criterion View of personal identity and forensic accountability, what Goldman wanted more than anything was some psychological continuer of Dorff to face the gallows for what he had done.

> So he planted in that psychological continuer of Dorff something that would incapacitate him from reverting back to G-D-G-D-G - call it integrity, pride, or what you will.

Is this an intelligible story? I think it is. I think conscience, for example, is something very much like this. Computer moles certainly are. Sophisticated viruses wire themselves to their hosts in such a way that, if an attempt is made to remove them, the whole system crashes. The same can be true of conscience. Dostoevsky wrote a book about it. [39]

Akin to conscience is reverence. I was once asked to spend a night in a synagogue to guard a cultural exhibit that was not due to open until the next day. There being nowhere else to sleep in the building, it was suggested to me I use a pew. There being nothing else to use as a pillow, I fetched a handful of prayer shawls. A decade after refusing to Bar Mitzvah - because, I pronounced, "It's all just primordial superstition!" - try as I might (and God knows I did) I could not force my head down onto those utterly indifferent strips of cloth!

Not all self-help books or videos are written or made by charlatans. There _are_ ways of escaping the unwelcome inhibitions and incapacitations of our upbringing - or, more generally, of our antecedence. The philosopher's fanciful pill is just one of them. But we have just supposed that, had there been available to Dorff an antidote to Goldman's curse, Goldman would have anticipated this and cursed him too with an allergy to the antidote. No doubt this is why Beatrice, recall, was disposed by the Loonies to "let it be known, in no less uncertain terms, that you don't wish to be deprogrammed."

But epiphenomenal frustration is not something to be wished only on our worst enemy. It is something, though we do not exactly wish on ourselves, we nonetheless do to ourselves, and do so knowingly. It is the price we willingly pay for the exhiliaration, or the consolation, of wanting things we simultaneously do not want to allow ourselves to have. This is precisely Ulysses' state after having commanded his men to tie him to the mast. It is the Disney World-goer's state halfway down Space Mountain. It is my own state half a day out on a bicycle tour. It is fate of Socrates, Jesus, Joan and More when offering themselves (respectively) to the judgment of the city, to the exigencies of human redemption, to the stake, or to the block.

It is what it is to be a part of something 'bigger than one's [current] self'.

9. The Prescriptivity Problem Revisited:

I am experiencing epiphenomenal frustration. Maybe I am the once-obedient but now-contemptuous Jew who is getting a crook in his neck because he cannot lay his head on "utterly indifferent scraps of cloth". Maybe I am a Cold War President bolting awake in a cold sweat from yet another of those recurring nightmares, the ones featuring a flock of geese and a full-scale nuclear response indifferent to my every order to stand down. Or maybe I am a concentration camp guard, standing naked and alone in one of the barracks as the handful of survivors limp to greet their liberators, staring helplessly at the garb on the floor that could save me from the gallows.

On MacIntosh's view, if I cannot change - and assuming I am rational - then the only explanation is I "really don't want to". On my view, I can say "I want to change, but something (or someone) else won't let me." So the disadvantage of MacIntosh's view is it seems grossly false to the phenomenon. And the disadvantage of mine is it courts the very counter-forensic compartmentalization of the self Fingarette rightly cautions us to avoid.

But on which view are we better equipped to exercise my self-monitoring module? Neither, it would seem. If I am so disposed I can indulge myself wondering, then investigating, and perhaps even discovering, why I am the way I am. No doubt I will find that someone - perhaps even some prior stage of myself - had an interest in making me this way, someone with interests other than, and not necessarily compatible with, my own.

Maybe it was the International Zionist Conspiracy that hoped to bind me to its political agenda by implanting in me this ridiculous, but insidious, fear of petty blasphemy.

True, I did support the Doomsday Device during the first months of my Presidency, when it was first put in place. But times have changed. The U.S.S.R. is no longer. Instead the geese have come back to Capistrano (del Norte). But the damn thing can't be turned off until 2023.

I'll bet it was that Jew that everyone said looked like me, that Goldman. I bragged to him once that if the Russians ever got close, I'd take his place and send the evidence up the smokestack. He looked at me long and hard. Put something in my brain with that look, he did. Sorcery, post-hypnotic suggestion, the black arts - Jews, they're good at those things. Haven't seen him lately. Must have died. Could even be his clothes on the floor I can't pick up. Damn him!

Or ... I know the only reason I won't feel fulfilled as a woman until I have children is that patriarchy's programmed me to feel this way. But, dammit, that's the way I feel!

One way out of this pickle is to say to MacIntosh, Well, so much the worse for its being a condition on defensibility that the psychology from which one self-projects be both first-ordered and current with respect to the self-projection! Or, failing that, then, So much the worse for defensibility!

Nonsense! scoffs the other voice within us. One need not go out of one's mind to get an angle from which to repair it. Look, some of us are more committed to our marriages than are others. But few of us, no matter how committed, have never once been tempted by a sexual indiscretion, and sorely tempted at that. Sexual arousal can be a powerful preference-altering pill. Consult your preferences in the throes of such a pre-coital passion. Maybe you will find there a preference not to prefer the continuation of this sexual encounter. But pre-coital passion leaves no room to remember what your now-absent spouse means to you. (That is not surprising, given the project that makes pre-coital passion so compelling.) That is, there is no time to act upon the preference to prefer to stop. What you need to do now is act upon the preference to stop. What you need to do now is act upon a preference you do not now have.

On MacIntosh's view, if you stopped, and if your doing so was rationally defensible, you could only have done so because you were able to effect, and therefore did have, the first-order preference to do so. On my view this is just _ad hoc_, _post facto_ question-begging and, phenomenologically, it is simply false. If you stopped, and if your doing so was rationally defensible, and since you did not have the first-order preference to do so, you could only have done so because it is both rational and possible to act contrary to current first-order preferences.

So which of these two views shall we adopt? Well, that all depends on which of the two projects we are pursuing. Are we trying to save the marriage? Then adopt mine; because using MacIntosh's circuitry, by the time I could convince myself to stop, the marriage would be over. Are we trying to satisfy current preferences? Then adopt his; because otherwise we will be forever enslaved by someone else's 'scripts'. But deciding between _those_ two projects can only be done from some _third_ perspective; and so on _ad infinitum_.

10. Personal Identity Revisited:

But there is a further difficulty with MacIntosh's position. The third objection, recall, was that even if Goldman preferred becoming what he himself would prefer did not come into being at all over ceasing to be entirely, it was a mistake for him to have replaced himself with G-D because by becoming G-D he did cease to be entirely. This is because "if R-relatedness is needed for identity, and preference changes" the likes of the (unincremented) transition from Goldman to G-D "violate R-relatedness" [40], G-D cannot be Goldman.

But, thinks MacIntosh, this is too quick. If "your values change," however thoroughly, "through rational reflection in an expression of your original values, the change serves them; they change through an activity characteristic of persons, the exercise of their rationality". [41] And so in what sense does the change violate R-relatedness? After all,

> R-relatedness is just relatedness of psychological states by processes that distinguish a psychology as such, and of these, rational changes of such states are paradigms ... [Furthermore, just as] one way for an agent's later beliefs to R-relate to his earlier ones is for the earlier to justify the later and to cause the later's formation because they justify them, [so] a similar causal/justificational relation can R-relate one's earlier and later desires. [And surely among] kinds of desire justification is: being [desires] the

> having of which can cause conditions earlier preferred. [42]

Thus, concludes MacIntosh, "B is the same person as A so far as B's psychological states rationally evolved from A's." [43] So, it would follow, G-D is Goldman, just in case Goldman had the preferences MIEU postulates he must have had had he been fully MIEU-rational.

But now, it seems, MacIntosh has a circularity problem to contend with. G-D is the same person as Goldman just in case G-D rationally evolved from Goldman. But a condition of G-D having rationally evolved from Goldman is that G-D is the same person as Goldman. Or, to focus on the complementary arc: a condition of G-D having rationally evolved from Goldman is that G-D be the same person as Goldman. But G-D is the same person as Goldman just in case G-D rationally evolved from Goldman.

To escape this circularity MacIntosh must either abandon the claim that the 'rationally-evolves-from' relation is sufficient for personal continuance, or else he must abandon the claim that personal continuance is a condition of rational self-projection. In either case objection (3) stands. And so in either case MIEU fails.

That it fails no worse than MEIU and MEIUCP, I have no doubt. But all this suggests is that some kind of meta-move - perhaps along the lines already suggested - might be in order.

11. Epistemology on a Need-to-Know-Only Basis:

Candidates for self-effacement and replacement include beliefs, convictions, capacities, and (quite possibly) ecological systems. They include desires, doubts, dispositions and descriptive theories. They include hopes, fears, memories and knowledge. And they include personalities, propositions, prescriptive theories and, perhaps, even publications. Accordingly, an enquiry into the logic of self-effacement promises important insights not only for Gauthier's project, i.e. meta-ethics and ethics, but also for epistemology and the philosophy of science, for the philosophy of mind and of the emotions, for the philosophy of law and of the environment, and for meta-philosophy. Here is a representative sampling, albeit a sampling of two.

The problem of skepticism arises from within an entirely human project - and probably a very specialized human project at that. So it is only as a certain kind of human that we want to 'know', in the sense of wanting what we call 'justified true belief'. Now by, "But why do we want to know?", I do not mean, "What are our reasons for wanting to know?", but rather, "What accounts for our wanting to know?" And the answer to that, I take it, is something along the lines suggested earlier, namely that:

> insofar as there were natural selective advantages to having an 'asking' function, but no _dis_advantages to not placing limits on that function, natural selection just never bothered to limit it.

So whereas Descartes' official answer to the problem of error in Meditation Four is that our will to know outstrips our means for doing so, he would have produced an answer more akin to the one provided above had he seen - as subsequent naturalized epistemologists have seen - that knowing in general can be given the same treatment as Descartes himself gives _sensing_ in Meditation Six. There he rightly observes that

> corporeal things may not all exist in a way that exactly corresponds with my sensory grasp of them. [44]

Indeed, he might have added, it may be that _no_ corporeal thing exists in a way that exactly corresponds with my sensory grasp of them. For what can it be for God to be no deceiver, and yet to allow that we are sometimes - as in the case of the phantom limb - _incorrigibly_ deceived, save that such deception is the result of

> the best system that could be devised ... [that] is most especially and most frequently conducive to the preservation of the healthy man. [45]

So if "the best system conducive to the preservation of the healthy man" was such "that no corporeal thing exists in a way that exactly corresponds with my sensory grasp of them," God would still be no deceiver. So what it is for X to be no deceiver of Y is just for X to conduce her (in this case creative) behaviour to the preservation of Y. So we humans can and do know - in the sense of 'knowing' that God's not deceving us entails. So all that remains for skepticism to mean is our not being able to know in some sense indexed to our purposes.

So, let us ask again, "Why do we want to know?" And this time we do mean, "For what reason do we want to know?" And now the only answer could be, "Oh, just idle curiosity!" If we were to then ask, "Why are we idly curious?", and since it would be a nonsensical question if by it we meant, "For what reason are we idly curious?", it follows we could only mean, "What accounts for our idle curiosity?" And for that question we do seem to have an answer. We are idly curious because we need to be generally curious, and because it is easier to make something indiscriminately curious than only selectively curious. If it should turn out that being idly curious does not "conduce to the preservation of the healthy man" - if, for example, there were a mass epidemic of Cartesian skeptics committing suicide over their epiphenomenal frustration at not being able to know - no doubt next time around God - or, as I prefer, natural selection - will install the extra circuitry to make us only selectively curious.

So, it turns out, Cartesian skepticism is just our residual epiphenomenal frustration in the wake of God - or, as I prefer, natural selection - having found no need to efface our idle curiosity.

12. Theodicy:

But, peenges the theodical skeptic, clearly God did efface our capacity to know. So even if

1) that capacity were neutral with respect to forwarding God's projects, and even though
2) God is no deceiver in the naturalized epistemological sense cited above, since
3) there are things we want to know and could know without jeopardizing Her Own benevolent projects,
4) She remains nonetheless malevolent.

Unless, of course, knowing what we otherwise could know would jeopardize Her benevolent projects. But then what could Her benevolent projects be such that our knowing would jeopardize them?

But hold. If they are benevolent in any non-trivial sense, they must likewise be our projects. We may have self-effaced them, but they must have been our projects 'originally'. So we can re-pose the question as, "What must our projects have been such that our knowing would jeopardize them, and so for the mediated self-effacement of the capacity to know we needed, solicited, and elicited God's assistance?"

Game-theory meets the exegesis of Genesis 2:15? Why not?

13. Might the Logic of Self-Effacement be Self-Effacing?

If candidates for self-effacement and replacement include descriptive theories and publications, it would be a dereliction of duty not to consider the possibility that this manuscript should be condemned to the pyre the moment it is completed. So let us consider it.

I once read a paper on the game-theoretic reduction of morality to a group of colleagues in which, to disarm what opposition I could by being self-deprecatory, I referred to my position as "the Repulsive Point of View" or RPV. And a year later I delivered an even more cold-blooded reduction to the same group, upgrading it to "the Consistently Repulsive Point of View", or CRPV. One of my listeners sent me a note expressing worries about the social consequences of such reductions ever catching on. And I replied with an undertaking never to publish in Scientific American.

I treat this problem with humour because, quite frankly, I am uneasy about it. So let us consider it - seriously.

Nothing that has been said so far can preclude the possibility that part or even all of our meta-ethical theorizing might be itself the object of an ethical judgment, and as such itself a candidate for self-effacement. That is, morality passes judgment on how we view morality. Suppose, then, that as a result of our meta-ethical theorizing we come to the conclusion - as I think we very well might - that it is morally wrong to hold the meta-ethical view from which we drew this self-same conclusion? Do we reject the theory, the judgment, both or neither? Certainly there are no logical grounds in any of this to re-think the theory, nor for that matter to regard as suspect the judgment passed on it. A meta-ethical claim cannot contradict an ethical one, nor can an ethical claim contradict a meta-ethical one. All they can do is refuse to share the same brain. Still, we are left - are we not? - in the unenviable position of having to reject a set of beliefs we believe to be true.

But perhaps morality is just one of those activities that require a certain lack of self-consciousness - not unlike dancing, bowling, or making love. Maybe we just develop a kind of intramental toggle-switch, akin to our duck/rabbit-ing in perception or, in taking action, our being now-determined/now-free.

Alternatively, we might be conceited enough to imagine ourselves among Plato's guardians; that we can handle certain truths that others cannot. True, philosophers are no less responsible than scientists for the injuries their findings inflict on the communities they serve. Nor can they take refuge any more than can scientists behind an ethic of indifference to the context of dissemination. Einstein lived to regret that indifference. So did many a social Darwinist. Even if "the truth you've spoken [is being] twisted by knaves to make a trap for fools" [46], who is the knave who serves these fools up?

Nor is Gauthierianism unique in this respect. Much of what counts as the philosophy of the environment today amounts to pointing out that the metaphysics of old is fouling our nest; and that maybe what we need to preserve it is a new one. And the same has been said more generally of something called 'modernity'. But virtually none of this literature actually attempts to refute the target doctrine. And this used to irritate me. Until, that is, I realized what they realized all along, namely that the truth or falseness of the doctrines in question were never themselves in question.

Is the logic of self-effacement self-effacing? I am not sure. Much less am I sure what I would do if social conscience ever dictated that I stop doing what it is I do and that I stop thinking about what it is I think about. Hopefully it will never come to that. But if it does, "x" marks the spot to put the match.

x

Notes to "Introduction and Roadmap" and to Chapter One

01. Gauthier, David, Morals by Agreement, Clarendon, Oxford, 1986

02. Ibid., pp. 15-16, 167-70, 177, 179-81, 285-7, 355

03. Ibid., pp. 14-16, 70n., 74, 145, 136-41, 143-6, 150, 154-8, 246-8, 265, 271, 304-5, 322, 340

04. Ibid., v, vi, 16, 192-3, 200-35, 255-69, 277-80, 290-94, 339-40, 342-4

05. The word 'holocaust' means burnt offering, which in turn suggests the Jews of Europe were sacrificed, either for their fellow Jews, for Christendom, or for humankind. Each of these claims strike many survivors and/or children of survivors as highly offensive. The word 'shoah', by contrast, simply means destruction, and is therefore theodically, and so acceptably, neutral.

06. For the seminal literature by H.L.A. Hart and Lon Fuller on just this issue, see Joel Feinberg's anthology, Philosophy of Law, Wadsworth, Belmont, 1991.

07. Locke's account of both synchronic and diachronic personal identity occupies the whole of Chapter 27 in the Second (i.e. 1694) Edition of his Essay Concerning Human Understanding.

08. Ibid., at the outset of Section 26

09. Ibid., Section 7

10. Ibid., Section 15

11. Perry, John, A Dialogue on Personal Identity and Immortality, Hackett, Indianapolis, 1978

12. Williams, Bernard, "The Self and the Future", The Philosophical Review, Vol. 79, No. 2, April, 1970

13. Parfit, Derek, Reasons and Persons, Oxford U.P., 1986

14. Butler, Joseph, "Of Personal Identity", being the first appendix to his The Analogy of Religion, 1736,

and anthologized in John Perry's *Personal Identity*, U. of C. Press, Berkeley, 1975

15. Reid, Thomas, "Of Identity" and "Of Mr. Locke's Account of Our Personal Identity", being Chapters Four and Six, respectively, of his essay "Of Memory" in his *Essays on the Intellectual Powers of Man*, 1785, and anthologized in John Perry's *Personal Identity*, U. of C. Press, Berkeley, 1975

16. Quinton, Anthony, "The Soul", *The Journal of Philosophy*, Vol. 59, No. 15, July, 1962

17. Grice, H.P., "Personal Identity", *Mind*, Vol. 50, October, 1941

18. Perry, John, "Personal Identity, Memory, and the Problem of Circularity, in his own anthology, *Personal Identity*, U. of C. Press, Berkeley, 1975

19. Parfit, op. cit.

20. Ibid., p. 262.

21. Parfit is vague about just *how* much similitude is enough for the Q-relation; but such Sorites problems are hardly peculiar to his theory.

22. Parfit, op. cit., p. 220

23. Ibid., pp. 282-287

24. Ibid., p. 220

25. As of this writing, the *very* latest, it seems, is Peter Unger's *Identity, Consciousness, and Value*, O.U.P., New York, 1990

26. Ibid., p. 258. For more detail see Lewis, David, "Survival and Identity", *The Identities of Persons*, A. Rorty (ed.), U. of C. Press, Berkeley, 1976.

27. Hirsch, Eli, *The Persistence of Objects*, University City Science Centre, Philadelphia, 1973, and *The Concept of Identity*, Oxford U.P., 1982

28. Kolak, Daniel and Martin, Raymond, "Personal Identity and Causality: Becoming Unglued", manuscript

29. Viminitz, Paul, On Dropping the Causal Connexity Requirement on Identity Claims Across Person-Stages, M.A. Thesis, Dalhousie University, 1988

30. Williams, Bernard, Problems of the Self, Cambridge U.P., 1973

31. Quine, W.V.O., "A Review of Identity and Individuation", Journal of Philosophy, 1972, p. 490. I make some further comments related to this issue at the outset of Chapter Eight.

32. Williams, Bernard, "The Self and the Future", The Philosophical Review, Vol. 79, No. 2, April, 1970

33. We will be returning directly to the issue of personal identity towards the end of Chapter Eight.

34. The motion picture was called Lady Hawke.

35. Nozick, Robert, Anarchy, State and Utopia, Basic, N.Y., 1974. See also The Examined Life, Simon and Schuster, N.Y., 1989, The Normative Theory of Individual Choice, Garland, N.Y., 1990, and The Nature of Rationality, Princeton U.P., 1993

36. Eliot, T.S., "The Naming of Cats"

37. this being the appeal of Superman,

38. and this the Garden of Eden.

39. Hobbes concurs. "And this is granted to be true by all men," says he, "in that they lead Criminals to Execution, and Prison, with armed men, notwithstanding that such Criminals have consented to the Law, by which they are condemned." (Leviathan p. 70)

40. Of course none of this is meant to settle the question of whether there are or are not principles worth giving one's life for. Nor do I see that here, at least, an answer is urgently called for. I will say that if there are any such principles, they are probably few and far between. Certainly the death of Socrates - to those of us only mildly plagued by Kantian intuitions - was less heroic than idiotic. On

the other hand it would be a nasty world indeed - and quite probably an uninhabitable one - were human beings such that their damn-the-torpedoes thresholds could at no point be breached.

So let me answer the question obliquely. I think we manufacture such thresholds. And I think we do so for reasons that are entirely natural. But they are by no means, on that account, any the less ennobling. (Apropos of which, see my comments on the advantages of MacIntosh's 'MIEU' in Chapter Eight.) For in affairs of the heart, no less than in social and political affairs, such thresholds are often all and precisely what keeps tyranny at bay!

Beyond this point - that is, to essay any less oblique an answer - there be dragons! But let us at least show our colours.

If, by something being 'of intrinsic value' we mean it can be _of_ value notwithstanding it neither is, was, or ever will be either currently or dispositionally valu_ed_, the very idea is unintelligible. That is, that 1) valuing is a diadic relation, I take to be analytic. Analytic too I take to be that 2) the valuing relation is not a transitive one. That is, we can, and often do, value what's valued by those we value. And more often than not we do so because they value them. But from this it does not follow we value them in virtue _simpliciter_ of their valuing them.

Now from (1) it might be supposed I think that 3) we cannot value what we're no longer extant to value. And from (2) it might be supposed I think that 4) neither can we value what only extant others can value. From which it would seem to follow that, on my view at least, 5) nothing can be worth dying for. To which I counter that, though (5) follows from (3) and (4), (3) does not follow from (1), nor does (4) from (2). For nothing in either (1) or (2) precludes our valuing things that are either no longer extant or as yet absent; nor is there anything in either (1) or (2) that blocks our valuing something that came, or will come, into a world entirely indifferent, or, for that matter, hostile, to it. (See the discussion in MacIntosh on this point rehearsed in Chapter Eight.)

So neither Dorff nor Goldman can be excused on the sole grounds that their lives were at stake. That one's life is at stake gives us _prima facie_ reason to exculpate otherwise indefensible behaviour; but it is never a conclusive reason.

41. Rawls, John, A Theory of Justice, Harvard U.P., 1971. Once again, I will have more to say about this in Chapter Eight.

42. I will have more to say on this 'offensiveness' when I review some of the consequences of MacIntosh's MIEU in Chapter Eight.

43. The view that homosexuality is nurtured rather than natured - and therefore can be reversed - is most famously espoused by Elizabeth Moberly in The Psychology of Self and Other, Tavistock, London, 1985, and in The Psychogenesis of the Early Development of Gender Identity, Routledge and K.Paul, 1993.

44. We will see in Chapter Eight what happens when we replace 'agendas' here with 'meta-agendas'.

45. I need not pronounce here on what those grounds might be. I have my own intuitions; and, for what they are worth, here they are. But whether they strike a responsive chord, or whether they seem just idiosyncratic, nothing much hangs in the balance.

    First of all, as with any crime, crimes against one's own identity - if such there be - are exculpated in direct proportion to the import of what one hopes to gain, in inverse proportion to the value of what she is prepared to lose, and in inverse proportion too to the direness of any further injury she might reasonably expect her actions to produce. So if what Goldman stood to gain by sacrificing his personal identity had been anything so obscene as an extra turn around the yard, the sexual use of another prisoner for an hour, or the privilege of going thirty-first rather than twenty-seventh into the gas chamber, or had he believed that rather than merely becoming Dorff in the epiphenomenal privacy of his own mind he would in fact be foisting on the world another Dorff-very-much-in-the-flesh, we would, I think, be much less sympathetic than we are.

    Second, I think it would be a cruel anachronism to attribute to Goldman anything approaching an adequate appreciation of the dimensions and historical significance of what was happening in those camps, and of why it was happpening, an appreciation that took even post-Shoah Jewry years, if not decades, to develop. By the same token I also think it would be mere caricature to imagine that the mind-set of the National Socialist is exhausted by its anti-Semitism. So I am not sure that, in terms of betraying their respective ethnicities, either man is more culpable than the other. In any event it seems to me a mistake to attribute much of anything very human, much less anything peculiarly Jewish, to a man who had been through what Goldman had.

By contrast, we have every reason to believe - or at least no reason not to believe - that in Dorff, notwithstanding he was in fact once Goldman, all the second-order virtues, like pride, courage, integrity, and so on, should have been fully in place and operational. Furthermore, he must have recognized (albeit falsely) that he was (albeit in his mind justifiably) himself the author of the circumstances that were now making it advisable for him to masquerade as Goldman. So he must have taken himself, however justifiably, to be escaping the consequences of his own actions. (Goldman, recall, far from trying to escape what was his own responsibility, rushed to embrace what was not.)

So maybe all our intuitions are telling us is just this. Cowardice in the face of evils authored by others is one thing. Cowardice in the face of self-authored evils is something else again. Is this not precisely what makes MacBeth's eleventh-hour "Lay on MacDuff!" so stirring and so redeeming of him in our eyes? Indeed, is this not precisely what makes Dorff's pre-trial resurrection a kind of vindication of the man's integrity, notwithstanding how monstrous we find the resurrected man to be?

This is not to suggest he necessarily ought not to have betrayed his Nazi values and principles in order to save his skin. This is a judgment only he could have made. It is to suggest, however, that for him - but not so for Goldman - sacrificing his skin for his values and principles was a live option, one upon which he deliberated and, apparently, one he rejected. So the least we are entitled to is to think less of his commitment to those values and principles. Which is just to say - and the word is the same in both Yiddish and German - we think him less the 'mensch' for it.

46. Of course by "having got it just about right" here I do not mean to suggest they were in the right. About that I am of several minds.

47. I consider the counter-position at the outset of Chapter Eight.

48. See the opus of Roger Shiner.

49. Leviathan, p. 63

50. Ibid., p. 62

51. Rorty, Richard, Objectivity, Relativism, and Truth, Cambridge University Press, 1991

Notes to Chapter Two

01. The Republic, 358a-359b

02. In Sub-Section 1 of Section 1 of Chapter 1 of the Principia Ethica, G.E. Moore notes that

> The hypothesis that disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole, may be most plainly seen to be incorrect by consideration of the fact that, whatever definition be offered, it may be always asked, with significance, of the complex so defined, whether it is itself good. To take, for instance, one of the more plausible, because one of the more complicated, of such proposed definitions, it may easily be thought, at first sight, that to be good may mean to be that which we desire to desire. Thus if we apply this definition to a particular instance and say 'When we think that A is good, we are thinking that A is one of the things which we desire to desire,' our proposition may seem quite plausible. But, if we carry the investigation further, and ask ourselves 'Is it good to desire to desire A?' it is apparent, on a little reflection, that this question is itself as intelligible, as the orginal question 'Is A good?' - that we are, in fact, now asking for exactly the same information about the desire to desire A, for which we formerly asked with regard to A itself. But it is also apparent that the meaning of this second question cannot be correctly analysed into 'Is the desire to desire A one of the things which we desire to desire?'; we have not before our minds anything so complicated as the question 'Do we desire to desire to desire to desire A?' Moreover any one

> can easily convince himself by inspection that the predicate of this proposition - 'good' - is positively different from the notion of 'desiring to desire' which enters into its subject; 'That we should desire to desire A is good' is not merely equivalent to 'That A should be good is good.' It may indeed be true that what we desire to desire is always also good; perhaps, even the converse may be true: but it is very doubtful whether this is the case, and the mere fact that we understand very well what is meant by doubting it, shews clearly that we have two different notions before our minds.

And, I think, Moore is quite right. But, it is to be noted, (what has since come to be called) the 'open question argument' for intra-level non-reductionism tells not at all against inter-level reduction.

03. See Rawls, John, A Theory of Justice, Harvard U.P., 1971

04. In Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem, Richmond Campbell reports (p. 4) that "The problem was first formulated about 1950 by a social psychologist, Merrill M. Flood, and an economist, Melvin Dresher." But its first appearance in the 'literature' was R.D. Luce and H. Raiffa, Games and Decisions, Wiley, New York, 1957. Luce and Raiffa (p. 94), in turn, attribute the story to a certain A.W. Tucker.

05. For a detailed analysis of the symmetrist argument, see Chapter Four.

06. The distinction between tuism and altruism is an important one for contractarians if they are to defend themselves against the charge - levelled especially but not exclusively by feminists - that contractarians cannot account for genuinely altruistic desires even in a state of nature. A tuistic desire is one with respect to a co-player's preferences. An altruistic desire is one with respect to another's interests.

07. Once again, the claim is not that natural selection *is* the explanation for why we have the dispositions we do, but rather and only that it would not be surprising if it were.

08. I have chosen the word 'deconstruction' here, but perhaps at my own peril. By it I do not mean whatever it has come to mean for practitioners of post-modernism in whatever the practice of post-modernism might be.

09. See Chapter Four.

10. The Logic of Leviathan, Clarendon, Oxford, 1969, is the name of David Gauthier's own first major work.

11. The compliance problem, in a nutshell is this. From the rationality of agreeing, committing, or even deciding to perform a certain act, it does not straightforwardly follow that it is therefore rational to follow through on that agreement, commitment or decision. Suppose, for example, Column and Row agree or commit or decide to institute as Sovereign some third party S. Of course S is instituted not by this agreement or commitment or decision alone - nor, come to think of it, by this agreement or commitment or decision at all! - but rather by the acts that arise from it, i.e., in this case, Column and Row disarming themselves. Now as Gregory Kavka points out - see Moral Paradoxes of Nuclear Deterrence, Cambridge U.P., 1987 - under normal circumstances,

> it is convenient, for many purposes, to treat a prior intention to perform an act as the beginning of the act itself. (p. 19)

But if Column becomes convinced Row either has or will disarm, any reason she had to agree or commit or decide to disarm in the first place has just disappeared. And, mutatis mutandis, similarly reasons Row. So to suppose one can get out of a PD is just to suppose one was never in one in the first place.

12. The Republic, 358e-359b

13. Leviathan, pp. 64-65

14. Ibid., p. 72

15. Ibid., p. 62

16. Huntington, Samuel P., The Soldier and the State, Belnap Press, Cambridge, 1972

17. The Republic, 375b-d

18. Ibid., 376a

19. The motion picture is called Butch Cassidy and the Sundance Kid.

20. Ackerman, Bruce, "Why Dialogue?"

21. In addition to Gauthier's The Logic of Leviathan, Clarendon, Oxford, 1969, an excellent game-theoretic reduction of Leviathan can be found in Jean Hampton's Hobbes and the Social Contract Tradition, Cambridge U.P., 1986.

22. Leviathan, p. 102

23. Ibid., p. 101

24. Ibid., p. 63

25. Ibid.

26. Ibid.

27. See Hampton, Jean, Hobbes and the Social Contract Tradition, Cambridge U.P., 1986.

28. Ibid.

29. Ibid.

30. Ibid.

31. The Republic, 359d-361d

32. For a discussion of those virtues of explanatory hypothesis-formation that have informed my thinking here, see Chapter 6 of W.V.O. Quine's and J.S. Ullian's The Web of Belief, Random House, 1970.

Notes to Chapter Three

01. The title of this chapter is a salutory play on the title of David Gauthier's own, The Logic of Leviathan, Clarendon, Oxford, 1969.

02. Gauthier, David, Morals by Agreement, Clarendon, Oxford, 1986

03. Ibid., p. 167

04. Ibid., p. 169

05. Ibid.

06. Ibid., pp. 15-16, 167-170, 177, 179-181, 285-287, 355

07. Once again, the claim is not that natural selection *is* the explanation for why we have the dispositions we do, but rather and only that it would not be surprising if it were.

08. Danielson, Peter, Artificial Morality, Routledge, New York, 1992

09. Ibid., p. 19

10. Ibid., p. 4

11. Ibid., p. 5

12. Ibid., p. 69

13. Ibid.

14. Gauthier (1986), op. cit., p. 174

15. Danielson, op. cit., p. 200

16. Ibid., p. 38

17. Ibid., p. 39

18. That AM does in fact constrain itself only by procedural possibility is what I shall be disputing momentarily.

19. Danielson, op. cit., p. 195

20. Ibid., p. 6

21. Ibid., p. 198

22. Ibid., p. 200

23. Ibid., pp. 154-5

24. Ibid., p. 65

25. Ibid., p. 89

26. Campbell, Richmond, "Gauthier's Theory of Morals by Agreement", The Philosophical Quarterly, Vol. 38, No. 152, 1988, p. 351

27. Danielson, op. cit., p. 85

28. Smith, Holly, "Deriving Morality from Rationality", Peter Vallentyne (ed.), Contractarianism and Rational Choice: Essays on Gauthier, Cambridge U.P., 1991, p. 242, n. 18

29. Danielson, op. cit., 86

30. Ibid., p. 82

31. Ibid., p. 76

32. Ibid., p. 83

33. Ibid., p. 86

34. Ibid., pp. 83-4

35. Ibid., p. 89

36. Ibid., p. 95

37. Ibid., p. 160

38. Ibid., p. 95

39. Ibid.

40. Ibid.

41. Ibid., pp. 95-6

42. Ibid., p. 95

43. Ibid., p. 96

44. Ibid., p. 93

45. Ibid., p. 160

46. Ibid., p. 161

47. Gauthier (1986), op. cit., p. 226

48. Danielson, op. cit., p. 161

49. Ibid., p. 39

50. Campbell, op. cit., p. 351

51. Kavka, Gregory, Moral Paradoxes of Nuclear Deterrence, Cambridge U.P., 1987, p. 19

52. Ibid., p. 24

53. Ibid., p. 25

54. Ibid., p. 27

55. Campbell, Richmond, in a talk delivered to the Canadian Philosophical Association at the Learneds in at Queens University in Kingston, Ontario, in May, 1991

56. Nielsen, Kai, Equality and Liberty: A Defense of Radical Egalitarianism, Rowman and Allenheld, Totowa, N.J., 1985

57. Rawls, John, A Theory of Justice, Harvard U.P., 1971

58. Nozick, Robert, Anarchy, State, and Utopia, Basic, N,Y,, 1974

59. Narveson, Jan, The Libertarian Idea, Temple U.P., Philadelphia, 1988

60. Gauthier (1986), op. cit., pp. 14, 16, 136-141, 143-146, 150, 155, 157-158, 268,

61. Ibid., pp. 178-179, 225-227, 230

Notes to Chapter Four

01. The title of this chapter is a play on the title of Lewis' paper (see #2 below) to which it is a response.

02. Lewis, David, "Prisoners' Dilemma is a Newcomb Problem", Philosophy and Public Affairs, Vol 8, No. 3, Spring 1979, pp. 235-240

03. Though Newcomb's Problem is reputed to have been invented by physicist William Newcomb in the early 1960's, it first entered the literature with Robert Nozick in 1969.

04. Leslie, John, "Ensuring Two Bird Deaths With One Throw", Mind, Vol. 100, No. 397, January 1991, pp. 73-86

05. Lewis, op. cit., pp. 235-236

06. Ibid., p. 236

07. Ibid., p. 237

08. Ibid., p. 238

09. Ibid., p. 239

10. Leslie, op. cit., p. 74

11. Lewis, op. cit., pp. 239-240

12. Leslie, op. cit., pp. 76-78

13. Ibid.

14. Ibid.

15. Ibid.

16. Davis, Lawrence, "Prisoners, Paradox, and Rationality", American Philosophical Quarterly, Vol. 14, No. 4, October 1977, pp. 319-327; reprinted in Campbell and Sowden, Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem, U.B.C. Press, Vancouver, 1985, pp. 45-59. All page references are to the latter.

17. Rapoport, Anatoly, Two-Person Game Theory, Michigan U.P., Ann Arbor, 1966, p. 141

18. Watkins, John, "Comment: Self-Interest and Morality", in S. Korner (ed.), Practical Reason, Blackwell, Oxford, 1974

19. Davis, in Campbell and Sowden (eds.), op. cit., p. 48

20. Sen, Amatya, "Reply to Comments", in S. Korner (ed.), op. cit., p. 80

21. Davis, in Campbell and Sowden (eds.), op. cit., p. 51

22. Leslie, op. cit.

23. Ibid.

24. Lewis, David, Counterfactuals, Blackwell, London, 1973
and, especially, On the Plurality of Worlds, Blackwell, Oxford, 1986

25. Kavka, Gregory, Moral Paradoxes of Nuclear Deterrence, Cambridge U.P., 1987, pp. 57-70

26. Once again, the claim is not that natural selection is the explanation for why we have the dispositions we do, but rather and only that it would not be surprising if it were.

27. Sobel, J. Howard, "Not Every Prisoner's Dilemma Is a Newcomb Problem", in Campbell and Sowden (eds), op. cit., pp. 263-274

28. I would like to thank J. Howard Sobel, Peter Danielson, Andrew Irvine, Leslie Burkholder and Louis Marinoff, for helping me think my way through this problem. But most of all I would like to thank Don Stewart who, though not himself a game-theorist, took a keen interest with me in the problem for no other reason, initially at least, than that I took a keen interest in it. Such collegiality is as invaluable as it is rare.

Notes to Chapter Five

01. Viminitz, Paul, "The Manifesto of the Silicon Valley Liberation Front - or - Might Machines be Morally Considerable?", manuscript

02. Dennett, Daniel, "Intentional Systems", in *Brainstorms*, Bradford, 1981, and *The Intentional Stance*, MIT Press, Cambridge, 1987

03. James, William, "What Pragmatism Means", in *Pragmatism*, Reynolds, 1907

04. Dworkin, Ronald, *A Matter of Principle*, Harvard U.P., 1985

05. Fuller, Lon, "Positivism and Fidelity to the Law - A Reply to Profesor Hart", 71 *Harvard Law Review*, 630, 1958

06. Hart, H.L.A., "Positivism and the Separation of Law and Morals", 71 *Harvard Law Review*, 593, 1958

07. Hobbes, Thomas, *Leviathan*, p. 61

08. Ibid.

09. Ibid., p. 60

10. Ibid., p. 61

11. Ibid., p. 63

12. Held, Virginia, "Non-contractual Society: A Feminist View", *Canadian Journal of Philosophy*, Supplementary Volume 13

13. Rawls, John, *A Theory of Justice*, Harvard U.P., 1971

14. Held, op. cit.

15. Sartre, Jean-Paul, *Being and Nothingness*, Washington Square, 1966, p. 10

16. Quine, W.V.O. and J.S. Ullian, *The Web of Belief*, Random House, 1970, pp. 64-82

17. See Chapter Eight.

18. Putnam, Hilary, Reason, Truth and History, Cambridge U.P., 1981

19. For a discussion of this see especially Williams, Bernard, Imagination and the Self, Oxford U.P., 1966, Moral Luck, Cambridge U.P., 1981, and Ethics and the Limits of Philosophy, Harvard U.P., 1985.

20. See Jaggar, Alison, Feminist Frameworks, McGraw-Hill, N.Y., 1993 and especially Feminist Politics and Human Nature, Rowman and Allenheld, Totowa, N.J., 1983.

21. Jaggar herself does not make this point.

22. Braybrooke, David, Meeting Needs, Princeton U.P., 1987

23. Braybrooke himself does not make this point.

24. Kipling, Rudyard, "If".

25. Goodman, Nelson, Fact, Fiction and Forecast, Bobbs-Merrill, Indianapolis, 1973

26. The sentiment, if not the words, are typically expressed in Gary Madison's The Logic of Liberty, Greenwood, N.Y., 1986.

27. Woodward, James, "Paternalism and Justification", in Wes Cragg (ed.), Contemporary Moral Issues, McGraw-Hill Ryerson, Toronto, 1987, pp. 249-264

28. Ibid., p. 250

29. Ibid., p. 258, 260

30. Ibid., p. 251

31. Ibid., p. 254-255

32. Ibid., p. 260

33. Ibid., p. 255

34. Ibid., p. 252

35. Ryle, Gilbert, The Concept of Mind, Barnes and Noble, N.Y., 1969

Notes to Chapter Six

01. Danielson, Peter, Artificial Morality, Routledge, London, 1992, p. 45

02. Ibid., p. 46. "It is a mistake," cautions Danielson, "to define morality too widely. Not every secondary rule specifying the egoistic first principle should count as a moral rule. Otherwise, we trivially incorporate all of economics, game and decision theory into morality."

03. The most effective conveyances of this point are made cinemagraphically by the final shoot-out scene in Sergio Leone's The Good, the Bad, and the Ugly, and by the first Russian Roulette scene in Michael Cimino's The Deer Hunter.

04. In a manuscript entitled "Morality and Prudence - Extending the Danielsonian Model", I argue that even on Danielson's own terms this kind of prudentiality can count as morality.

05. See note # 03.

06. Just as I am a non-literalist with respect to biblical exegesis, I have incorporated this advise with adjustments for inflation. He also told me once, "Buy a gun and know how to use it!", but he has since denied ever having said it. Still, the lesson to be learned is - be careful what you say to your children. They might be listening!

07. Danielson, op. cit., p. 93

08. See Chapter Eight.

09. Once again, here I am following Quine in Chapter 6 of his and Ulian's The Web of Belief, Random House, New York, 1970.

Notes to Chapter Seven

01. Danielson, Peter, Artificial Morality, Routledge, London, 1992, p. 14

02. For a remarkably unguarded articulation of revolutionary theory, see Guevara, Ernesto, Guerrilla Warfare, Monthly Review Press, N.Y., 1961 or his Che Guevara on Guerrilla Warfare, Praeger, N.Y., 1961

03. Mavrodes, George, "Conventions and the Morality of War", in Beitz et al (eds.), International Ethics, Princeton U.P., 1985, pp. 75-89. This is why the National Socialist war criminals who took the fewest pains to cover their tracks were those who had most convinced themselves their Reich would last a thousand years. This may be also why the Shoah could only have happened in a country as civilized as Germany. Anywhere else the Nuremberg Laws would have more pursuasively put their victims on notice. This is why too the Israelis are as bellicose as they are. "Never again!" is an idle boast, since one's own actions alone seldom guarantee the desired outcome. "Never again!" is shorthand for "Never again trust!" and "Seldom if ever appease!"

04. See my discussion of MacIntosh's discussion of this in Chapter Eight.

05. Danielson, op. cit., p. 161

06. Parfit, Derek, Reasons and Persons, Clarendon, Oxford, 1984

07. See my discussion of MacIntosh's discussion of this in Chapter Eight.

08. See my discussion of MacIntosh's discussion of this in Chapter Eight.

09. Of course one has to be careful about identifying which positions are stronger and/or weaker than which. If, for example, I was an atheist only because I was first and foremost an impossibilist, then by my strongly self-effacing my atheism I would have to intend only my becoming a possibilist. But as to

what impossibilism's *weak* fallback position might be, God alone knows! That said, what is meant by doxastic 'well-orderedness' or 'compatibility' is notoriously problematic. Do doxastic operators aggregate, do they decompose, do they penetrate? And what about deontic operators embedded *within* doxastic operators? Do *they* aggregate, decompose and penetrate? Do they practicate? And what about doxastically embedded *modal* operators? Would that I could, I can do nothing here other than to bracket these questions; and to acknowledge that my account remains, therefore, woefully incomplete.

10. Danielson, op. cit., p. 160

11. Gauthier, David, *Morals by Agreement*, Clarendon, Oxford, 1986, p. 269

12. Ibid.

13. Rawls, John, *A Theory of Justice*, Harvard U.P., 1971

14. For a taste of the complexities involved in understanding counterfactuals, see the material cited in note #4-24.

15. Demos, Raphael, "Lying to Oneself", *Journal of Philosophy*, Vol. 57, 1960, pp. 588-595

16. Canfield, John and McNally, Patrick, "Paradoxes of Self-Deception", *Analysis*, Vol. 21, 1961, pp. 140-144

17. Canfield, John and Gustavson, Don, "Self-Deception", *Analysis*, Vol. 23, 1962, pp. 32-36

18. See the corpus of Roland Puccetti.

19. Fingarette, Herbert, *Self-Deception*, Routledge & Kegan Paul, London, 1969

20. Ibid.

Notes to Chapter Eight

01. Leslie, John, Value and Existence, Blackwell, Oxford, 1979

02. The assumption that what matters in survival just *is* personal identity has been very much out of vogue since the early 1970's.

03. Rawls, John, A Theory of Justice, Harvard U.P., 1971

04. MacIntosh, Duncan, "Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values", Journal of Philosophy, Vol. XC, No. 4, April, 1993, pp. 163-180. See also MacIntosh's "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma", Pacific Philosophical Quarterly, LXXI, 1991, pp. 9-32, and his "Preference-Revision and the Paradoxes of Instrumental Rationality", Canadian Journal of Philosophy, XXII, 1993, pp. 503-30.

05. Ibid., p. 163

06. Kavka, Gregory, Moral Paradoxes of Nuclear Deterrence, Cambridge U.P., 1987

07. MacIntosh (1993), op. cit., p. 164

08. Ibid.

09. Ibid.

10. Ibid., p. 165

11. Ibid.

12. Ibid.

13. Ibid., pp. 165-6

14. Ibid., p. 166

15. Ibid.

16. Ibid., p. 165

17. Ibid., p. 166

18. Ibid.

19. Ibid., p. 167

20. Ibid., p. 168

21. Ibid., p. 167

22. Ibid., p. 168

23. Ibid., p. 169

24. Ibid., p. 170

25. Ibid., pp. 171-2

26. Ibid., p. 172

27. Ibid.

28. Ibid., p. 173

29. Ibid., p. 174

30. Ibid.

31. Ibid.

32. Ibid.

33. Ibid., p. 175

34. Ibid.

35. Ibid.

36. Ibid., p. 176

37. See notes # 44 & 45

38. That we can only in vain reflect upon our having no choice about engaging in this reflection - the burden of our own freedom, and all that! - I take to just about the sum total of the entire corpus of Twentieth Century continental thought, so far as I can penetrate it. But enough about the limits of my own powers of penetration.

39. Crime and Punishment

40. MacIntosh (1993), op. cit., p. 178

41. Ibid.

42. Ibid. p. 179

43. Ibid. p. 180

44. Cottingham, John, (trans.), Descartes, Rene Meditations on First Philosophy, Cambridge U.P., 1986, p. 55

45. Ibid., p. 60

46. from Rudyard Kipling's "If".

Bibliography

01. Axelrod, Robert (1984), The Evolution of Cooperation, Basic, New York

02. Beitz et al (eds.) (1985), International Ethics, Princeton University Press

03. Braybrooke, David (1987), Meeting Needs, Princeton University Press

04. Campbell, R. and Sowden, L. (eds.) (1985), Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem, University of British Columbia Press, Vancouver

05. Cragg, Wes (ed.) (1987), Contemporary Moral Issues, McGraw-Hill Ryerson, Toronto

06. Danielson, Peter (1993), Artificial Morality, Routledge, New York

07. Dennett, Daniel (1981), Brainstorms, Bradford

08. Dennett, Daniel (1987), The Intentional Stance, MIT Press, Cambridge

09. Elster, Jon (1984), Ulysses and the Sirens: Studies in Rationality and Irrationality, Cambridge University Press

10. Elster, Jon (ed.) (1985), Rational Choice, New York University Press

11. Elster, Jon and Hylland, Aanund (eds.) (1986), Foundations of Social Choice Theory, Cambridge University Press

12. Fingarette, Herbert (1969), Self-Deception, Routledge & Kegan Paul, London

13. Gauthier, David (1969), The Logic of Leviathan, Clarendon, Oxford

14. Gauthier, David (1986), Morals by Agreement, Clarendon, Oxford

15. Hampton, Jean (1986), Hobbes and the Social Contract Tradition, Cambridge University Press

16. Hirsch, Eli (1973), The Persistence of Objects, University City Science Centre, Philadelphia

17. Hirsch, Eli (1982), The Concept of Identity, Oxford University Press

18. Hobbes, Thomas, Leviathan

19. Jaggar, Alison (1983), Feminist Politics and Human Nature, Rowman and Allenhead, Totowa, N.J.

20. Jaggar, Alison (1993), Feminist Frameworks, McGraw-Hill, New York

21. Kavka, Gregory (1986), Hobbesian Moral and Political Theory, Princeton University Press

22. Kavka, Gregory (1987), Moral Paradoxes of Nuclear Deterrence, Cambridge University Press

23. Korner, S. (ed.) (1974), Practical Reason, Blackwell, Oxford

24. Leslie, John (1979), Value and Existence, Blackwell, Oxford

25. Luce, R.D. and Raiffa, H. (1957), Games and Decisions, Wiley, New York

26. Madison, Gary (1986), The Logic of Liberty, Greenwood, New York

27. Narveson, Jan (1988), The Libertarian Idea, Temple University Press, Philadelphia

28. Nielsen, Kai (1985), Equality and Liberty: A Defense of Radical Egalitarianism, Rowman and Allenheld, Totowa, N.J.

29. Nozick, Robert (1974), Anarchy, State, and Utopia, Basic, New York

30. Nozick, Robert (1981), Philosophical Explanations, Harvard University Press

31. Nozick, Robert (1989), The Examined Life, Simon and Schuster, New York

32. Nozick, Robert (1990), The Normative Theory of Individual Choice, Garland, New York

33. Nozick, Robert (1993), The Nature of Rationality, Princeton University Press

34. Parfit, Derek (1986), Reasons and Persons, Oxford University Press

35. Perry, John (ed.) (1975), Personal Identity, University of California Press, Berkeley

36. Plato, Republic

37. Quine, W.V.O and Ullian, J.S. (1970), The Web of Belief, Random House, New York

38. Rapoport, Anatol (1966), Two-Person Game Theory, Michigan University Press, Ann Arbor

39. Rapoport et al (1976), The 2 X 2 Game, University of Michigan Press, Ann Arbor

40. Rawls, John (1971), A Theory of Justice, Harvard University Press

41. Rorty, A. (ed.) (1976), The Identities of Persons, University of California Press, Berkeley

42. Vallentyne, Peter (ed.) (1991), Contractarianism and Rational Choice: Essays on Gauthier, Cambridge University Press

43. Williams, Bernard (1966), Imagination and the Self, Oxford University Press

44. Williams, Bernard (1973), Problems of the Self, Cambridge University Press

45. Williams, Bernard (1981), Moral Luck, Cambridge University Press

46. Williams, Bernard (1985), Ethics and the Limits of Philosophy, Harvard University Press